Predicting Age from Microbiome Data: Benchmarking Multi-Source Machine Learning Methods

University of Turku
Department of Computing
Master of Science (Tech) Thesis
Data Science
December 2024
Shadman Ishraq

Supervisors:
Prof. Leo Lahti, DSc
Tuomas Borman, MSc (tech)

UNIVERSITY OF TURKU
Department of Computing

Shadman Ishraq: Predicting Age from Microbiome Data: Benchmarking Multi-Source Machine Learning Methods

Master of Science (Tech) Thesis, 56 p., 2 app. p.
Data Science
December 2024

The microbiome holds significant potential as a predictor of biological processes, including age, due to its dynamic interaction with human health. This study addressed the challenge of predicting age using microbiome data by benchmarking tree-based machine learning models such as Random Forest (RF), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGBoost), in addition to the IntegratedLearner method. In this study, the LifeLines DEEP dataset was utilized, incorporating relative abundance, marker abundance, and pathway abundance data to predict age. Both single-omic and multi-omics models were developed, focusing on evaluating the impact of data integration on predictive performance. The results demonstrated that multi-omics models outperformed single-omic models, with GBM trained on multi-omics data sets and the stacked model used by the IntegratedLearner method achieved the highest predictive accuracy. Functional data sets, particularly pathway abundance, exhibited stronger correlations with age compared to taxonomic dataset, underscoring their significance for age prediction. Despite challenges posed by sparse, zero-inflated data and limited microbial diversity, the findings suggest that multi-omics integration enhances model performance and provides valuable insights into age-related biological processes.

Keywords: microbiome, gut microbiota, multi-omics, single-omic, IntegratedLearner, GBM, RF, XGBoost

Contents

List of Figures

List of Tables

1 Introduction

Microbes are microscopic organisms that inhabit the human body for varying periods of time, playing critical roles in health, disease, and overall human physiology. Recent studies have highlighted the uniqueness of the human microbiome and recognized the influence of microbes residing in the body as a potential biomarker for diverse applications in medicine and healthcare. For instance, microbial patterns in the human body have been found to be as distinctive as fingerprints, enabling possibilities for human identification and personalized medical interventions [1]. However, the exploration of these patterns for predictive modeling remains a computationally demanding task, requiring the application of advanced data science methodologies.

The field of microbiome research has undergone a rapid transformation, driven by technological advancements in genomics and bioinformatics. Early approaches, such as 16S rRNA gene sequencing, provided taxonomic insights into microbial communities, while whole-genome shotgun sequencing enabled more comprehensive functional analyses. Recent methodologies like integrated multi-omics analysis, single-cell multi-omics, and spatial transcriptomics have further expanded the scope, offering exceptional resolution to microbial ecosystems and their interactions with the host [2]. However, these advancements have also led to the generation of high-dimensional, sparse, and compositional data sets, which present unique challenges for data preprocessing, feature selection, and predictive modeling.

From a data science perspective, microbiome data sets are particularly intrigu-

ing. Their compositional nature indicates that the features (e.g., relative abundance of microbial taxa) are expressed as proportions of a whole, making traditional statistical methods unsuitable without appropriate transformations. Furthermore, the high sparsity of the data due to the absence of certain microbes in many samples intensifies the difficulty of analysis. These characteristics demand specialized computational approaches, such as log-ratio transformations, dimensionality reduction techniques, and advanced machine-learning models capable of handling complex data structures.

Despite the challenges, microbiome data offers significant opportunities for predictive modeling. One promising application is the prediction of human age, which involves identifying patterns in microbial communities that correlate with the aging process. While individual omics data sets, such as relative abundance, marker abundance, and pathway abundance, have been utilized for specific applications, research comparing the predictive capabilities of these data sets within a unified framework is limited. Additionally, the integration of multi-omics data into predictive modeling using tree-based machine learning models and the IntegratedLearner method remains insufficiently explored, particularly in contexts such as age prediction. Interestingly, the ability to predict age using a sparse and compositional dataset like relative abundance data alone highlights the compelling potential of microbial omics in advancing predictive modeling efforts.

This thesis addresses these research challenges by systematically evaluating the predictive capabilities of various microbial omics data sets and their integration for age prediction. Employing advanced machine learning techniques and comprehensive biological analyses, this work aims to provide deeper insights into the relationship between the human microbiome and aging while contributing to developing more accurate microbiome-based predictive models. Overall, this thesis highlights the convergence of microbiome research and data science, emphasizing the role of

computational methodologies in addressing biological challenges.

1.1 Research Objectives

The focus of this thesis is to explore the potential of microbial omics data sets for predicting age and to compare the predictive performance of various data sets in this context. The study utilizes species relative abundance data from the LifeLines DEEP dataset and a multi-omics dataset that combines relative abundance, marker abundance, and pathway abundance data. A comprehensive approach is adopted to analyze the data, beginning with descriptive, taxonomic, and correlation analyses to gain deeper insights into the relationships between microbial features and age. The effectiveness of these features is then evaluated in predictive models using advanced machine learning techniques, including Random Forest(RF), Extreme Gradient Boosting (XGBoost), and Gradient Boosting Machine (GBM), along with the IntegratedLearner method. By assessing the predictive performance across data sets and analyzing the underlying biological patterns, this research aims to provide an understanding of the relationship between the human microbiome and age while contributing to the development of more accurate models for microbiome-based age prediction.

1.2 Thesis Structure

The thesis begins with an exploration of the microbiome and its significance in understanding human health, providing a theoretical foundation for the study. It inspects the prior research on the relationship between microbiome composition and age, identifying gaps that motivate the present work. The discussion then extends to the packages, tools, and methodologies utilized in microbial community analysis, such as curatedMetagenomicData, mia, and caret, along with advanced techniques

like Indicator Species Analysis, Shannon Diversity, Beta Diversity, and methods for feature importance analysis.

The analysis begins with an overview of the LifeLines Deep dataset, presenting insights into the structure and its relevance for this thesis. Data exploration follows descriptive, taxonomic, and diversity analyses to uncover patterns and characteristics of the microbiome data. Indicator Species Analysis and correlation analysis are employed to identify relationships within the data, providing a deeper understanding of the dynamics between microbiome features and age.

Building on this groundwork, the modeling workflow outlines the process of preparing and preprocessing the data for machine learning models. Techniques are developed to predict age using models such as RF, GBM, and XGBoost. Model performance is evaluated using metrics like R^2 and MAE, while model-based feature ranking highlights the most influential microbial features in predicting age. The results of the modeling efforts are then presented, focusing on the accuracy and reliability of predictions and the identification of significant features. These findings are examined based on the random baseline and the significance of the model's performance.

Finally, the thesis discusses and interprets the results in relation to the research objectives, compares the performance of different models, examines the implications of the findings for microbiome research and age prediction, and highlights the limitations of the study while suggesting future directions for research to improve microbiome-based predictive models.

2 Literature Review

2.1 Microbiome Research

A microbe is an organism which is minuscule (smaller than 100 μm) to the unaided human eye and periodically it can be a single cell, a cluster of cells, or some of the microbes can also be multicellular. Microbes can be a wide variety of organisms, including fungi, bacteria, algae, protozoa, plants such as green algae, and small animals such as rotifers and planarians, and some scientists even consider viruses as non-living microbes [3]. Human bodies contain trillions of microbes, ten times more than the number of human cells. Microbes are present in almost every part of the human body, including the skin, stomach and the nose [4]. Between the years 1665 and 1678, the first to formulate the concept of microbes were Robert Hooke and Antoni van Leeuwenhoek [5].

The term microbiome is formulated from the words "microbe" and "biome" referring to the complete biotic environment like residents, ecosystem, and genes. Whipps and his colleagues introduced the term "microbiome" in 1988 to characterise how a group of microbes interacts in a certain environment [6]. A human microbiome is consistently linked with the phrases "microbiome" and "microbiata". Fundamentally, "microbiota" refers to a taxonomy of the microbes associated with humans, while the term "microbiome" is the collection of those organisms and their genes [7]. Microbes living within the human body show significant influences on

human physiology [8]. These influences can be both favourable and harmful. It makes up a diverse and unique ecosystem that adapts seamlessly to the environmental conditions of each habitat across various body regions, including the skin, mucous membranes, intestinal tract, respiratory system, urogenital tract, and even the udders [8].

The human microbiome constantly changes state in response to host factors such as age, nutrition, lifestyle, hormonal changes, inherited genes, and underlying diseases [9]. The human gut is colonized by microbes from birth. The intestines of an infant is considered sterile or containing minimal microbes, but the colonization of microbes occurs shortly after the child birth, exposed from the mother's vagina [10]. It is assumed that the number of microbial cells in the human microbiota varies between 10 trillion and 100 trillion. This estimation is obtained from the total number of bacterial cells found in the colon, the largest intestine, which has the highest bacteria concentration of any body organ and containing approximately 38 trillion bacteria [11]. The colon is a part of the Gastrointestinal (GI) tract, which encircles a big part of the human body. A huge number of Bacteria, Archaea, Eukarya, and Viruses inhabit in the gut and among them 98% of the bacterial phyla are Firmicutes, Bacteroides, Proteobacteria, and Actinobacteria [12].The gut of the human body can be considered as the main reservoir of the microbiome.

The gut microbiota has numerous functional effects on human health and psychology. For instance, gut microbiota is referred as the "second brain", which controls anxiety, emotion, cognition, and pain through the gut-brain axis (GBA) [13]. The gut-brain axis acts as a dynamic communication network, bridging the brain's emotional and cognitive functions with the digestive system, which integrates the Central Nervous System (CNS) with the Enteric Nervous System (ENS) through a combination of biochemical signals and physical interactions.[14]. The CNS acts as the body's control center, by gathering and processing information from all regions,

and regulating a broad spectrum of activities to maintain the proper functioning of the organisms. [15]. However, the ENS is a complex network of neurons existing exclusively within the GI tract, playing a vital role in regulating and controlling key functions of GI physiology [16]. The interaction between the gut microbiota and the ENS is influenced by the Autonomic Nervous System (ANS), which regulates the essential functions of the GI tract and builds a link between the gut and the brain by triggering the responses of the nervous system. [17].

Most of the health disorders, and problems in the digestive system, such as allergies, obesity and disorders related to the central CNS, are found to be related to dysbiosis (an imbalance in bacterial composition) of the gut microbiota [18]. The gut microbiome is essential to health and disease through several mechanisms, most of which are associated with immune functionality, metabolic processes, and the protection against pathogens (infectious agents or germs). The relationship between the intestinal microbiota, the epithelial lining of the intestines, and the mucosal immune system is a highly intricate and dynamic process, acknowledging the fact that 70-80% of the body's immune cells reside in the gut [19]. The interaction between the intestinal microbiome and the host's immune system plays a pivotal role in the development of normal gut and systemic immune responses. Thus, the disruption may lead to a spectrum of diseases ranging from gastrointestinal to systemic immune-mediated illness [20]. Moreover, about 20% of all cancers are related to dysbiosis of the gut microbiota. The healthy gut microbiota (commensal microbiota) helps in the activation of inflammasomes, proteins that protect the health and integrity of host's cells and gut, so an imbalance in the gut microbiota can lead to tumor development or even cancer [21]. Additionally, the gut microbiota is capable of generating resistant carbohydrates that support intestinal barrier strength, metabolism, immune regulation, and blood pressure [22]. Therefore, a healthy gut microbiota is essential in maintaining proper immune functionality, supporting overall metabolic

health, and providing adequate protection against infections of all types. On the other hand, it may severely aggravate many diseases if the balance is disrupted.

2.1.1 Previous Studies on Microbiome and Age

Recently, numerous studies were conducted regarding microbiome and age. Studies showed that the human microbiome constantly changes as the age progresses. Studies like gut-age clock (gAge) and chronological age predictions established that a person's age can be predicted effectively based on their microbial abundance. The results of the gut-age clock (gAge) study revealed that there are strong association between gut microbiome, and human health profiles and particular microbial markers can represent an individual's health status, frailty level, and health condition [23]. Moreover, a study by the American Society of Microbiology described that it is possible to predict an individual's age within an average of four years solely based on the microbes of skin [24].

Starting from early adulthood, the diversity and the advancement of the human microbiome start to evolve, and the stability of the development declines until 65. The shrink of the microbiome dynamics is more noticeable after the age of 80 [25]. Based on various taxonomic compositional studies, Akkermansia was one of the most consistent abundances among the older population. While, bacteria like Faecalibacterium, Bacteroidaceae, and Lachnospiraceae reduced relatively with age [26]. Another study showed that in phylum level distribution Firmicutes, Bacteroidetes, and Actinobacteria exhibited consistency in both the young and old age groups. While, Euryarchaeota, Synergistetes, and Proteobacteria had higher consistency among centenarians and Bacteroidetes had higher consistency among the young age group [27].

The gut microbiome is linked with various diseases related to age, such as, exposure to various health conditions, disruption of the immune system, weakness, type-2

diabetes (insulin antagonism), cancer, and Alzheimer's [26] [12]. The microbial imbalance or dysbiosis occurring after a certain age can be stated as a significant reason for these types of diseases. Moreover, certain species like Akkermansia muciniphila have a crucial role in maintaining the virtue of the intestines and reducing inflammation. So, the loss of these functional species, along with some useful genera like Bifidobacterium and Faecalibacterium can raise the risk of inflammation and chronic diseases inside the body [28]. The biological changes occurring due to age make a surrounding change inside the human body (e.g. the GI tract), which increases the growth of microbial pathogens causing the decline of beneficial microbes [29].

Numerous factors of aging can cause an imbalance in microbial composition. Diet plays a crucial role in maintaining the development of healthy microbes. For older adults often the dietary plan changes from healthy food to low-fiber, high-sugar, fat food, which significantly influences the composition of gut microbiota [30]. Moreover, due to different health conditions aged people might have to consume numerous medications. Medication (e.g. antibiotics, inhibitors) is another reason that can transform the diversity of microbes, leading to dysbiosis [31]. As age progresses humans tend to lead lives inactively and the body undergoes immunosenescence (declined immune system), causing the vulnerability to uphold microbiome balance [30].

2.2 Machine Learning Models

Predicting continuous variables, such as age, from microbiome data requires the use of machine learning approaches that can effectively address the distinct challenges inherent in microbiome data sets, including high dimensionality, sparsity, and compositionality among features. Tree-based models, including RF, GBM, and XGBoost, are among the most commonly used methods for this purpose. For instance, there are studies that have predicted chronological age from microbiome data across

various body sites using Random Forest, while gradient boosting methods, such as GBM and LightGBM, have also been employed to predict chronological age from gut microbiome data, demonstrating their strong performance [24] [32]. Similarly, XGBoost has been applied to predict age using both gut microbiota and urine metabolites, showing strong predictive accuracy [33]. These studies collectively underscore the robustness and versatility of tree-based models in addressing the complexities of microbiome data and highlight their significant potential for advancing predictive research in this domain.

In addition to tree-based models, deep-learning approaches have been applied to analyze highly complex microbiome data sets, leveraging their ability to capture intricate patterns and interactions. However, these models typically require careful hyperparameter tuning and larger datasets to mitigate the risk of overfitting, which can be a limitation in microbiome research where data availability may be constrained. For example, a study has demonstrated the potential of host-based deep neural networks (DNN) for age prediction using gut microbiota, highlighting the ability of these models for comprehensive analysis [34].

Furthermore, multi-source learning approaches, like the IntegratedLearner framework, have been explored for combining predictions from individual models trained on separate omics data sets, such as metabolites, biomarkers, and species abundances. By integrating these diverse data sources, methods like IntegratedLearner have improved predictive performance and revealed novel insights into microbial ecosystems and their relationships with host phenotypes [35]. These advancements underline the growing importance of robust and versatile ML models in microbiome research, which continues to contribute significantly to the understanding of human health and disease.

2.3 Microbiome Data Science

Data science plays a crucial role in discovering the complex interactions between microbial communities and their environments by combining biology with computational tools. This section provides a description of the key resources and tools in microbiome data analysis, including numerous Bioconductor packages, Indicator Species Analysis methood, Shannon diversity, Beta diversity, and an overview of the curatedMetagenomicData package and the LifeLines DEEP dataset. Additionally, it highlights microbiome-specific data science frameworks and discusses the integration of machine learning techniques to derive actionable insights from complex microbial datasets.

2.3.1 R and Bioconductor

R is recognized as a crucial programming language for bioinformatics. It is compatible with analyzing complex microbiome data sets, applying statistical methods and testing, machine learning modeling, and data visualizations. Numerous tools and packages are used in R to cater the requirements of bioinformatics, which makes it the most widely used programing language in this field. R and Bioconductor are powerful tools in the field of bioinformatics and computational biology, working in combination to provide robust solutions for analyzing microbiome data.

Bioconductor [36] is an open-source project built on R that enhances bioinformatics capabilities by providing a curated collection of packages designed for the analysis and interpretation of genomic, transcriptomic, and microbiome data. Key packages such as mia [37] and phyloseq [38] are widely used for extracting valuable insights from microbiome datasets, supporting comprehensive and reproducible data analysis. Additionally, Bioconductor facilitates genome-scale analysis of high-throughput data, integration with biological metadata from databases like GenBank, PubMed, and curatedMetagenomicData [39] and offers powerful visualization tools

for data exploration and visualization.

2.3.2 Available Databases

There are numerous microbiome data sets available for research and practical uses, such as the curatedMetagenomicData package, which provides manually curated, regulized human microbiome data sets [39]. The dataset used for this thesis was collected from the curatedMetagenomicData package. Curated metagenomic data package provides either TSV files or well documented TreeSummarExperiment objects, including human-processed metadata . The data sets in the curatedMetagenomicData package use whole-metagenome shotgun sequencing instead of 16S rRNA gene sequencing, which offers much broader coverage of microbial communities by capturing genetic material from bacteria, fungi, archaea, and viruses [39]. 16S rRNA sequencing is the general approach for studying bacterial species and microbial diversity. While, this technique has a limited number of applications, because other types of taxa, such as viral or fungal, are not considered in this technique and it does not contain high-quality functional information [40].

Each of the datasets from curatedMetagenomicData package includes six different categories of dataset, including relative abundance, marker abundance, and marker presence produced by MetaPhlAn3 and gene families, pathway coverage, and pathway abundance produced by HUMAnN3 with the UniRef90 database [39]. Moreover, these data sets contain detailed sample metadata across various categories, such as health condition, age, gender, nationality and BMI. Each of these data types provides unique insight into the microbiome, puting together a more complete and complex picture of the microbiomes.

LifeLines DEEP Dataset

In curatedMetagenomicData package [39], there are 93 microbiome data sets in total, and this thesis will focus on the LifeLines DEEP dataset [41]. LifeLines DEEP dataset is a Dutch population based study comprising stool specimens collected from 1,179 individuals, with 44 samples excluded due to low read counts [42]. Similar to any other dataset of the curatedMetagenomicData package, the LifeLines DEEP dataset contains all the taxonomic and functional profiling data types. This thesis will focus on the species, pathways, and marker-specific abundance data sets of the LifeLines DEEP study. In total, these data types contain more than one hundred thousand omics information. Considering each data type, 646 predictive features are identified in the relative abundance data, 23,085 in pathway abundance data, and 78,698 in marker abundance data, which makes it well-suited for conducting multi-omics microbiome analysis. Moreover, the sample metadata of this dataset contains information like age, BMI, antibiotic usage, gender, age-category and disease condition.

2.3.3 Microbiome Data Science Frameworks

The mia package

The mia package [37] provides tools to explore microbiome data sets, including visualization, data simulation, summarization, community index estimation, and time series analysis, with integration into related packages like miaViz, miaSim, and miaTime. Moreover, this package contains various data sets of different microbiome studies as TreeSummarizedExperiment objects. By identifying key features, such as dominant phyla, and adjusting numerous parameters, users can gain insights into microbial ecology and community composition. The package also supports numerous data transformations and visualization tools for abundance across taxonomic

ranks or sample groups. Additionally, it enables the estimation of species prevalence based on taxonomic ranks and sample types. These capabilities facilitate a deeper understanding of microbiome dynamics, allowing researchers to explore variations across conditions, identify microbial patterns, and refine data interpretation for more accurate ecological assessments.

Diversity and Indicator Species

In 1972, Robert Harding Whittaker introduced the concepts of alpha, beta, and gamma diversity. According to Whittaker's foundational model, alpha diversity measures the variety of species within a single habitat, beta diversity measures the difference in species composition between different habitats, and gamma diversity measures the overall variety of species across multiple habitats in a specific area [43].

Alpha diversity measures species richness, evenness, and overall diversity, with certain indices also providing insights into species dominance. Depending on different aspects, numerous indices are utilized to assess the alpha diversity. For example, commonly used indices to assess alpha diversity include the Shannon Diversity Index, which considers both species richness and evenness; the Simpson Index, which measures the likelihood that two randomly chosen individuals belong to different species; the Chao1 and Fisher's indices, which estimate total species richness, including unobserved species; and the Berger-Parker Index and Simpson Dominance Index (Inverse Simpson Index), which quantify the dominance of the most abundant species. [44].

On the other hand, the fundamental index for estimating beta diversity is Whittaker's Beta diversity index, which is defined as the ratio of total species richness (Alpha Diversity) to the mean species richness across samples (Gamma Diversity) [43]. Similar to alpha diversity, beta diversity uses various indices to measure differences between species samples. Commonly used beta diversity indices include the

Jaccard dissimilarity index, which calculates the proportion of species absent from one sample; the Sorensen dissimilarity index, which also considers absent species but gives more weight to shared ones; the Bray-Curtis dissimilarity index, which compares the abundance of shared microbes between samples; and the Euclidean dissimilarity index, which measures the straight-line distance between samples in multidimensional space based on species abundances or relative proportions. [45] [46].

In addition to these diversity measures, Indicator Species Analysis (ISA) [47], developed by Dufrêne and Legendre, is a statistical method used in ecology to assess the relationship between species and specific groups or clusters (e.g., age categories). [48]. ISA is performed by calculating the indicator values, which are the product of "specificity" and "fidelity". Whereas, specificity is the indicator that compares the quantity of the species in a certain group and fidelity calculates the prevalence of the species in a certain group comparing to the other available groups [49]. Moreover, in this method permutation test is conducted to check the statistical significance of these relationships [48].

Specificity,

$$A_{ij} = \frac{\bar{x}_{ij}}{\sum_j \bar{x}_{i\cdot}}$$

Fidelity,

$$B_{ij} = \frac{n_{ij}}{n_{\cdot j}}$$

Indicator Value,

$$IV_{ij} = A_{ij} \times B_{ij}$$

Where,

- $\bar{x}_{ij}$;mean abundance or presence of certain species i in a group j.
- $\sum_j \bar{x}_{i\cdot}$;sum of the mean abundance or presence of certain species i in all the

groups j.

- n_{ij} ;number of samples in group j inhabited by species i.
- $n_{.j}$;total samples in group j.

[49]

Correlation Analysis

Correlation analysis is an essential technique for identifying key features within a dataset. Methods such as Pearson correlation, Spearman correlation, and Kendall's Tau are used to assess the relationships between features and the target variable. For instance, in a microbiome study, functional features that show a strong correlation with the target variable can be prioritized for modeling, while less relevant features, such as taxonomical ones with low correlation, may be excluded. This approach helps to filter less important features, reducing dimensionality and enhancing model performance. Furthermore, using correlation analysis as a preprocessing step ensures a more focused dataset, which is especially valuable for tasks that require efficient feature selection and interpretation.

2.3.4 Machine Learning Frameworks

Machine Learning is a powerful approach in microbiome analysis that leverages computational algorithms to extract meaningful insights from complex microbial community data. Numerous machine learning algorithms like logistic regression [50], random forest[51], support vector machines [52], lasso regression [53],XGBoost [54] and gradient boosting machines[55] can be used in microbiome analysis to uncover the functional relationships between microbial communities and the demographic variables. Machine learning frameworks are used for a structured approach in training, and deploying machine learning models. There are numerous frameworks that

offer in-built functions, algorithms, and utilities that simplify the development process of the models. In this research, the caret and IntegratedLearner packages were utilized to streamline model training and integration, providing a robust foundation for the analysis.

The caret package

The caret package [56] is a model-building and evaluating package for classification and regression tasks. It contains numerous models like random forest[51] ,XGBoost [54] and gradient boosting machines [55] . The caret package directly gives access to features like data splitting, data pre-processing, feature selection, feature importance, model tuning, parallel processing, and visualization [57]. The main purposes for using the caret package in this research were, the simple workflow of machine learning models, effective tools to use cross-validation in hyperparameter tuning, showcasing the features that impacted the model predictions significantly, and variation of in-built model evaluation on train set. The caret package along with the specific algorithm (XGBoost, GBM, RF) packages makes it more flexible to customize the hyperparameters to enhance the performance of the models.

The IntegratedLearner package

Similar to the caret package, IntegratedLearner [35] is also an open-source R package that serves as a unified machine learning framework for multi-omics prediction, utilizing data from both longitudinal and cross-sectional multi-omics studies. IntegratedLearner method generates results for each omics layer through a two-stage learning approach that begins with Bayesian additive regression trees (BART) [58] as base learners for each omics layer, followed by a meta-learning approach in the second stage to evaluate the weights of the layers based on the unseen data predictions from the first stage. Along with the BART, it also supports multiple machine

learning algorithms like Linear Mixed Effects models, Non-negative Least Squares, and Rank Loss Minimization for both base and meta-learners. In addition, the IntegratedLearner package incorporates over 50 machine learning algorithms from the SuperLearner package, along with several data analysis utilities, offering a wide range of model selection options within a robust estimation framework [35].

3 Methods

In this research, a comprehensive model pipeline was developed to predict the age using two data sets, the microbial relative abundance dataset, and the concatenated dataset combining relative abundance, marker abundance, and the pathway abundance. This section provides a detailed explanation of the pipeline and describes the key stages, including data preprocessing, model training, hyperparameter optimization, and performance evaluation.

3.1 Data Preparation

The microbiome data sets were prepared after obtaining the TreeSummarizedExperiment (TSE) objects from the curatedMetagenomicData package, which provided a standardized and structured format for microbiome data analysis. The taxonomic relative abundance data, along with the pathway and marker abundance data sets, were extracted from the TSE object as assay-type data and converted into a dataframe for analysis.

Given the compositional nature of the relative abundance dataset, a Centered Log Ratio (CLR) [59] transformation was applied as a preprocessing step. CLR transformation addresses compositional data challenges by managing compositionality and high sparsity, ensuring the analysis focuses on relative differences between components, with a pseudocount applied to zero values to enable log transformation while preserving component ratios [60]. For the pathway abundance and marker

abundance data sets, a log10 transformation was applied. This transformation handled highly skewed distributions by compressing large values and expanding smaller values, effectively standardizing the distributions [60]. These data transformation processes were integral in facilitating improved model training and ensuring a balanced contribution of features during the machine learning process. Preparing three distinct data sets, this process ensured compatibility with predictive modeling and allowed for meaningful comparative analysis.

For this study, two types of data sets were utilized for the machine learning models, single-omic data, comprising only microbial relative abundance data, and multi-omics data, combining relative abundance, pathway abundance, and marker abundance data sets. The single-omic dataset contained microbial relative abundance data for 646 features across 1135 samples. A filtering step based on the coefficient of variation (CV) was employed for the relative abundance dataset to reduce the dimensionality and retain the informative features of the dataset. To focus on features with meaningful variability, a CV threshold of 0.1 was applied. By conducting this preprocessing step, the number of features significantly reduced from 646 to 91, retaining only the most variable and informative features were retained for subsequent analysis.

For the multi-omics dataset, the relative abundance, pathway abundance, and marker abundance data sets were filtered to retain only the most relevant features for modeling, with the original data sets containing 646 features for relative abundance, 23,085 for pathway abundance, and 78,698 for marker abundance. Unlike the single-omic data, coefficient of variation (CV) filtering was not applied to the multi-omics data sets because it retained a large number of features, which would make the process computationally expensive. Instead, the Spearman correlation was applied to each dataset, with a correlation threshold of 0.16 to select the features that showed meaningful correlations with the target variable (age). After this process

the number of the features were reduced to 547, resulting in a substantial reduction in dimensionality. The three filtered data sets were then combined to create the multi-omics dataset.

Moreover, the IntegratedLearner model integrates data differently by creating three data sets: a feature table containing concatenated multi-omics features, with features as rows and samples as columns; a sample metadata table, which includes sample-specific metadata with columns for unique identifiers and target variables; and a feature metadata table, which contains feature-specific metadata along with unique identifiers and feature types [35]. Considering the heterogeneous nature of the data, the preprocessed multi-omics dataset was selected for the IntegratedLearner method and integrated accordingly to the model.

3.2 Development of the Models

In microbiome analysis research, tree-based models stand out for their significant performance, offering powerful insights and accurate predictions. [61]. In this research, three different tree-based machine learning models were devised to conduct a comprehensive performance analysis for the age prediction. To achieve this, regression models, such as RF, GBM, and XGBoost were selected. These models are well-suited for analyzing microbial omics data, which is often high-dimensional and sparse. By aggregating weak learners, tree-based models enhance predictive accuracy and reduce overfitting. Moreover, all of these models possess the capability to rank feature importance, offering valuable insights into critical predictors, which is key for microbiome analysis. Hence, these attributes made RF, GBM, and XGBoost particularly well-suited for this research.

The data sets (single-omic and multi-omics) utilized in this study, were split into 80% training split and 20% for testing split. For the further validation of the model's robustness, a 10-fold cross-validation was implemented, which split the

training set into 10 smaller sets. In each iteration, 9 folds were used to train the model while the remaining fold was used to test it. This process was repeated for each model, ensuring a comprehensive evaluation across different subsets of the data and enabling the selection of the best hyperparameter combination based on cross-validated performance metrics.

Building on this cross-validation approach, the RF model was implemented using the training set, configured with 400 trees to ensure a diverse ensemble capable of capturing complex patterns in the data. Hyperparameter tuning was performed by adjusting the number of variables considered at each split to optimize the model performance. In parallel, the XGBoost model was trained using a hyperparameter grid that included parameters such as the number of boosting rounds, maximum tree depth, learning rate, and regularization factors, all of which were fine-tuned to enhance the boosting process. For comparison, the GBM model was also trained with specific tuning parameters that controlled tree depth, the number of trees, the learning rate, and the minimum number of observations in a node, ensuring a fair comparison of model performance across different algorithms. All the three models (RF, XGBoost, and GBM) were trained on both single-omic and multi-omics data sets.

On the other hand, the IntegratedLearner method was employed exclusively for the multi-omics dataset. IntegratedLearner represents a new generation of integrative model designed to incorporate multiple omics data sets or layers. The IntegratedLearner model was developed by combining information from multiple sources, including features, sample characteristics, and metadata about the features. The model utilized Bayesian Additive Regression Trees (BART) [58] as the base learner and Non-Negative Least Squares (NNLS) [62] as the meta learner (both sourced from the SuperLearner package) to capture relationships within the data and make predictions. The model was trained using cross-validation, dividing the dataset into

folds to iteratively evaluate and enhance its performance, ensuring robustness in its predictions. After training, all tree-based models and the IntegratedLearner model were evaluated on the test set to assess their performance, allowing for a comparison of the predictive power of the models across each dataset for age prediction.

3.3 Evaluation and Model-based Feature Ranking

The evaluation process involves 10-fold cross-validation to assess the performance of the regression models for predicting age. The data is divided into 10 folds, and in each iteration, one fold serves as the test set while the model is trained on the remaining nine folds and generates predictions for the test fold. The Mean Absolute Error (MAE) and R-squared (R^2) values are calculated for each fold and averaged across all folds to evaluate the model's performance on the test set. Additionally, the actual and predicted values are compared to further assess the model's predictive performance. This approach provides a robust evaluation by considering variations in predictions across different subsets of the test data.

Furthermore, feature ranking reveals the role of individual features in enhancing the predictive performance of the machine learning models. Analyzing the importance scores generated by the models' built-in mechanisms allows for identifying features with the greatest influence on age prediction. In this thesis, the native functions of the machine learning models RF, GBM, and XGBoost were utilized to determine feature importance based on their contribution to the predictive accuracy of the models.

4 Results

This chapter presents the findings of the study, containing both data exploration and model performance results. The data exploration results include a descriptive analysis of the dataset, providing an overview of the target variable, features, and their distributions. Taxonomic analysis highlights the relative abundance and prevalence of microbial taxa. Additionally, the age category-based analysis is discussed, featuring the Indicator Species Analysis, which identifies taxa strongly associated with specific age groups, and diversity analysis, which includes Shannon Diversity and Beta Diversity, focusing on metrics that capture microbial community richness and evenness.

Moreover, this chapter also describes the evaluation of the performance of the machine learning models, detailing key metrics used to assess their effectiveness. A statistical comparison of the models is conducted to determine the significance of differences in their performance. Furthermore, the analysis explores the important features identified by each model, shedding light on their contribution to predicting the target variable. The aim is to provide a comprehensive assessment of model performance while identifying the features that most significantly influence predictions.

4.1 Descriptive Analysis Results

This section contains a brief description of the demographics of participants in the LifeLines DEEP study. In total, the dataset comprises 1,135 participants, where 58%

(661 individuals) of the population were female, and 42% (474 individuals) of the population were male. Among the participants, there were 92% (1,040 individuals) adults , 1% were school-aged children, and rest (7.4%) of the population were seniors. LifeLines DEEP dataset involved a wide range of samples, age varying from 18 to 81 years. The interquartile range (IQR) of the age was 34 to 54, showcasing a wide range of the middle-age population group. The median age of the population was 45 and the median value of the BMI was 24.6 kg/m^2. Most of the population participated in this study had a normal to slightly overweight body weight, with a BMI interquartile range (IQR) of 22.4 to 27.2 kg/m^2.

Summary of the LifeLines DEEP Dataset	
Characteristic	**Value**
Age (Median, IQR)	45 (34, 54)
Gender	
Female	661 (58%)
Male	474 (42%)
BMI (Median, IQR)	24.6 (22.4, 27.2)
Age Category	
Adult	1,040 (92%)
School-age	11 (1.0%)
Senior	84 (7.4%)
Country	Netherlands

Table 4.1: Summary of the dataset, including the percentage (%), and median (IQR) for continuous variables such as age, BMI, and age category distribution.

The Figure 4.1 depicted the age distribution of the samples of the LifeLines DEEP study. The age distribution of the study showed skewness, with a standard daviation of 13.6 and a mean age of 45.07 years (pointed by the red dotted lines).

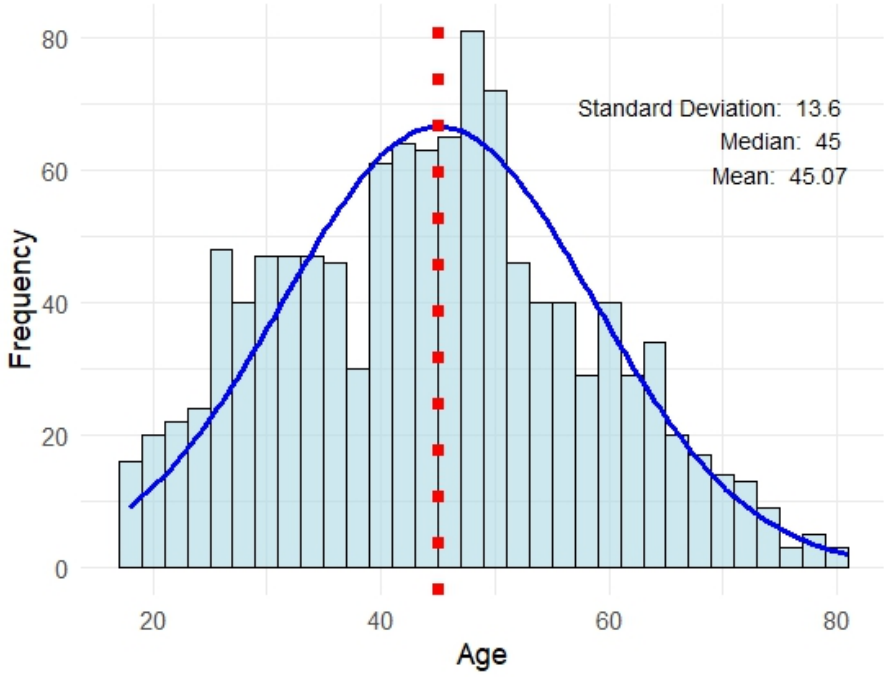


Figure 4.1: Age Distribution of the LifeLines DEEP Dataset.

Figure 4.2 depicted the distributions of features from the Pathway Abundance, Marker Abundance, and Relative Abundance data sets, each represented by three histograms. Across all datasets, the features exhibited right-skewed distributions, with most values concentrated near zero and a few outliers showed higher abundances. This pattern is characteristic of microbiome data, where certain taxa, markers, or pathways were rare and present in trace amounts, while others dominated in a subset of samples. The Relative Abundance dataset, in particular, reflected the compositional nature of microbiome data, where proportions were constrained and sparse. The observed skewness in all data sets highlighted the need for appropriate data transformations, such as logarithmic or CLR, to normalize the distributions and enhance the interpretability of downstream analyses. This visualization highlighted the challenges in analyzing microbiome data sets and the importance of preprocessing, to address sparsity and compositionality.

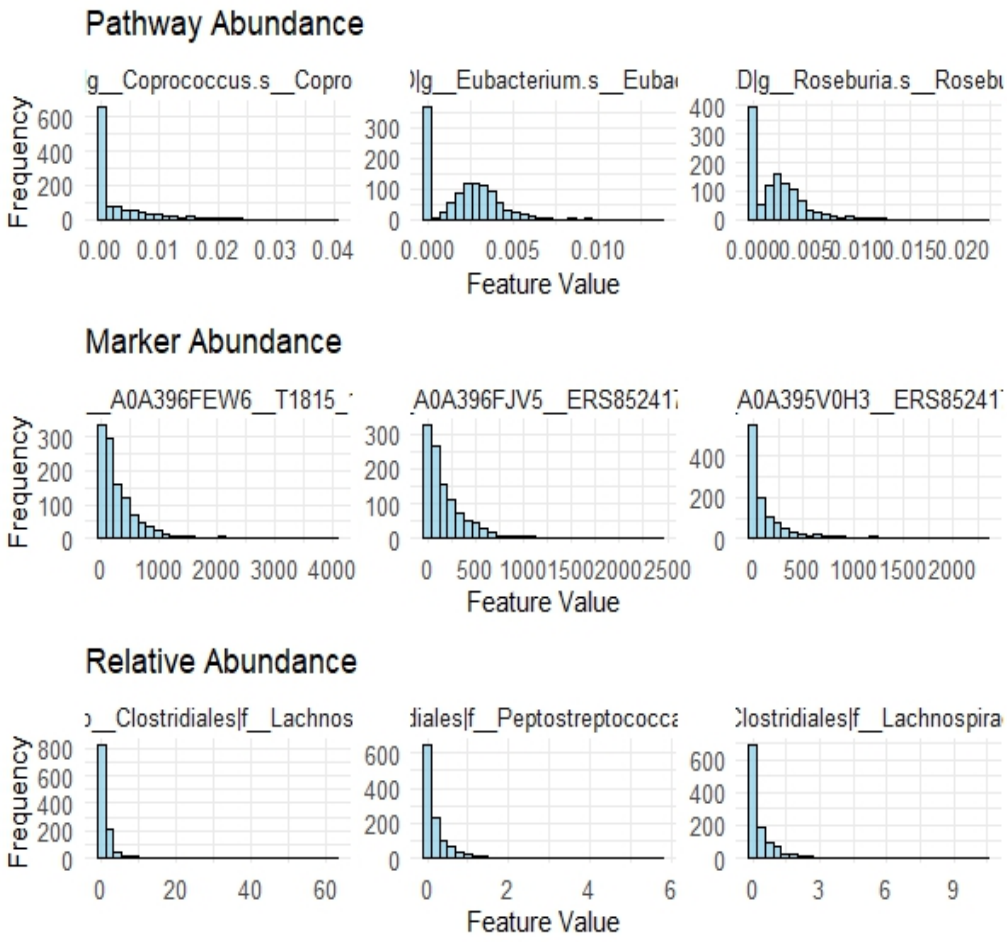


Figure 4.2: Distribution of selected features from the Pathway Abundance, Marker Abundance, and Relative Abundance datasets.

4.2 Taxonomic Analysis Results

In the LifeLines DEEP dataset, the relative abundance data contained information about 646 species. These species can be divided into three main groups of microbes or microorganisms, which can be classified into 13 taxonomic phyla. The majority of the species identified in this dataset were prokaryotic microbes like, Bacteria (638), and Archaea (6), and the remaining species can be classified as eukaryotes, which are more complex organisms than prokaryotes.

Identifying the dominant phyla and comparing the microbial composition helps

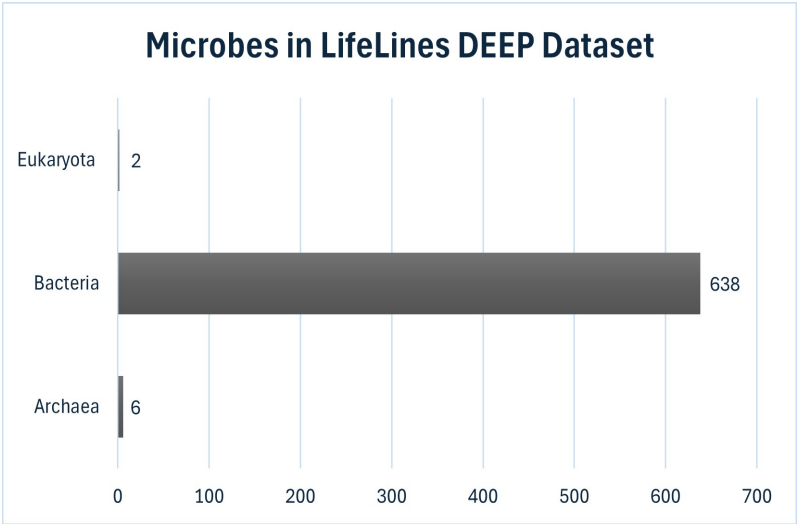


Figure 4.3: Different microbial groups in the LifeLines DEEP Dataset.

to understand the microbial diversity of the samples. Figure 4.4 depicted the microbial phyla composition over the collected samples based on their relative abundance. Overall, 13 phyla were identified in the LifeLines DEEP dataset, out of which Firmicutes were the most abundant (60.5%) phylum. Along with Firmicutes, Actinobacteria showed good consistency over the samples. In contrast, Euryarchaeota, Verrucomicrobia, and the rest of the phyla had shown low relative abundance.

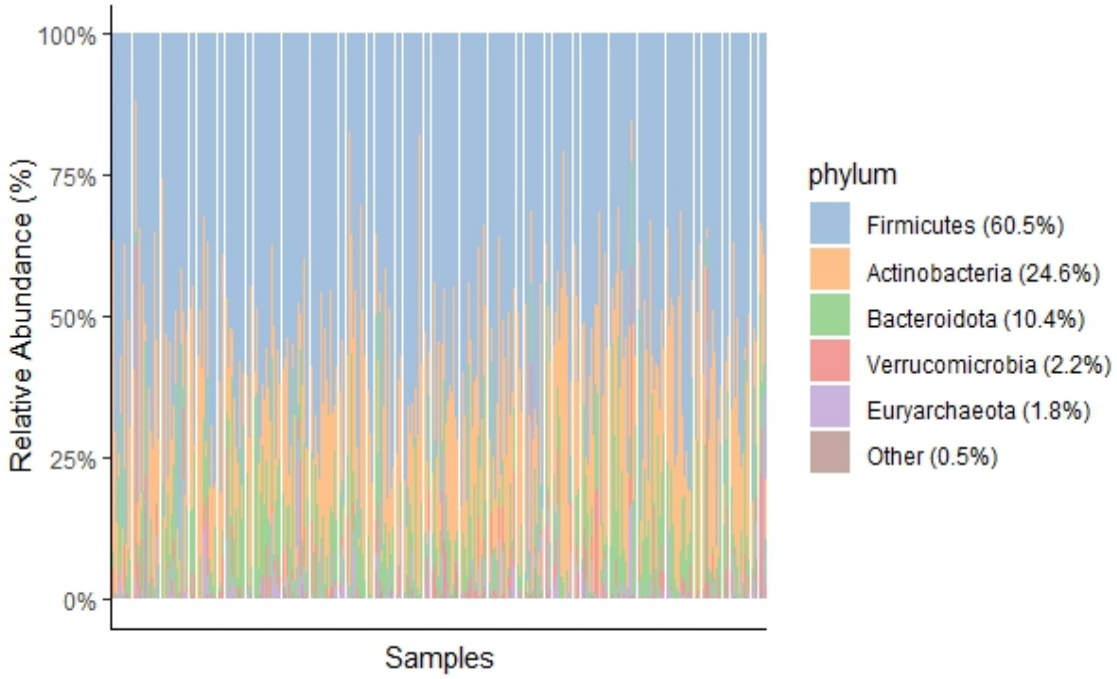


Figure 4.4: Top phyla based on relative abundance across the samples from the LifeLines DEEP dataset.

Figure 4.5 depicted the top prevalence of microbial species across samples. Faecalibacterium prausnitzii had the highest prevalence; the presence was more than 93% across samples. Where Collinsella aerofaciens and Dorea longicatena had similar prevalence around 87%. Other species, namely Eubacterium hallii, Coprococcus comes, and Bifidobacterium longum, had about 75% presence across the samples. While species like Akkermansia muciniphila, Methanobrevibacter smithii, and Bacteroides vulgatus showed much lower prevalence rates across the samples. Notably, among the top 20 species, the top three showed a prevalence of more than 85%, and the bottom three had a prevalence of less than 35%. So, it can be stated that the distribution of the microbial species underlines the scarcity of the microbiome dataset and there are a significant number of species that had a lower presence over the samples.

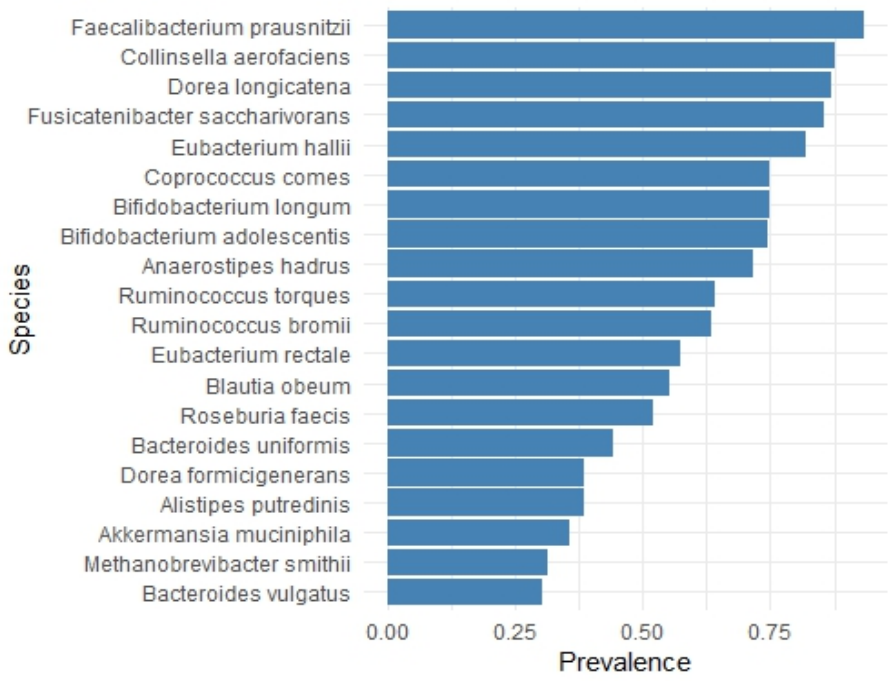


Figure 4.5: Top species based on prevalence across samples from the LifeLines DEEP dataset.

4.3 Indicator Species Analysis Results

Indicator Species Analysis can estimate the changes in the composition of the microbial community based on age category, offering valuable insights into the relationship between microbiome diversity and human health. By identifying specific microbial associations unique to different age groups, such analyses provide a foundation for understanding age-specific physiological or dietary factors that shape the microbiome.

This approach revealed distinct microbial associations for age categories such as adults, school-age individuals, and seniors. For adults, Eubacterium eligens emerged as the most strongly associated species, indicating its potential role in middle-aged individuals with an indicator value of 0.336 and a p-value of 0.037. Additionally, species such as Blautia obeum and Slackia isoflavoniconvertens were identified as significant contributors to the adult microbiome. Notably, Gordonibacter pame-

laeae displayed high significance in this group, underscoring its relevance to this age category. These findings highlighted the distinct microbial composition in adults, which may reflect age-specific physiological or dietary factors (Table 4.2).

For the school-age population, Gordonibacter pamelaeae, Eggerthella lenta, and Bacteroides fragilis demonstrated the strongest associations, characterized by indicator values surpassing 0.5 and highly significant p-values below 0.005, suggesting their prominence during the developmental years. Other species, including Blautia obeum and Actinomyces sp. ICM47, also showed significant associations, suggesting a unique microbial profile characteristic at this stage of life (Table 4.3).

Among seniors, species such as Slackia isoflavoniconvertens, Oscillibacter sp. CAG 241, Eubacterium eligens, Actinomyces oris, and Methanobrevibacter smithii were dominant, reflecting a diverse microbial community. The elevated indicator values of these species suggested their potential role in age-related biological processes in the senior population (Table 4.4).

Several species were identified across multiple age categories, demonstrating distinct variations in their indicator values. For instance, Slackia isoflavoniconvertens and Actinomyces oris were found in both adults and seniors, showing stronger associations in the senior group. Similarly, Gordonibacter pamelaeae appeared in both adults and school-age individuals but was more prominent in the school-age population. Blautia obeum, on the other hand, was present across all age groups, reflecting its versatile potential functionality.

Top Species among the Adult Population		
Species	**Indicator Value**	**P Value**
Eubacterium eligens	0.336	0.037
Blautia obeum	0.256	0.030
Slackia isoflavoniconvertens	0.254	0.010
Oscillibacter sp. CAG 241	0.249	0.014
Methanobrevibacter smithii	0.217	0.038
Gordonibacter pamelaeae	0.197	0.002
Clostridium disporicum	0.131	0.033
Actinomyces oris	0.113	0.017
Denitrobacterium detoxificans	0.098	0.024
Actinomyces sp. ICM47	0.092	0.049

Table 4.2: Indicator Species Analysis for the adult population based on their indicator values and statistical significance (p-values). While p-values, computed using the in-built permutation tests.

Top Species among the School Age Population		
Species	**Indicator Value**	**P Value**
Gordonibacter pamelaeae	0.610	0.002
Eggerthella lenta	0.598	0.003
Bacteroides fragilis	0.502	0.004
Blautia obeum	0.451	0.030
Actinomyces sp. ICM47	0.365	0.049
Bacteroides clarus	0.291	0.012
Actinobaculum sp. oral taxon 83	0.234	0.017
Streptococcus sanguinis	0.233	0.033
Streptococcus mitis	0.221	0.035
Escherichia albertii	0.181	0.001

Table 4.3: Indicator Species Analysis for the school-age population based on their indicator values and statistical significance (p-values). While p-values, computed using the in-built permutation tests.

Top Species among the Senior Population		
Species	**Indicator Value**	**P Value**
Slackia isoflavoniconvertens	0.553	0.010
Oscillibacter sp. CAG 241	0.543	0.014
Eubacterium eligens	0.445	0.037
Actinomyces oris	0.401	0.017
Methanobrevibacter smithii	0.360	0.038
Clostridium disporicum	0.359	0.033
Streptococcus gordonii	0.344	0.002
Denitrobacterium detoxificans	0.312	0.024
Mogibacterium diversum	0.278	0.037
Blautia obeum	0.277	0.030

Table 4.4: Indicator Species Analysis for the senior population based on their indicator values and statistical significance (p-values). While p-values, computed using the in-built permutation tests.

4.4 Diversity Analysis Results

Figure 4.6 depicted the distribution of the Shannon Diversity Index across three age categories adult, school-age, and senior. The Shannon diversity index is a measure that accounts for both species richness and evenness within microbial communities. The Shannon diversity index increased from the school-age group to the senior group. Whereas, the adult group showed intermediate diversity. The adult and school-age groups showed a lack of significance. Hence, these age categories have similar levels of microbial diversity. The LifeLines DEEP study includes individuals aged from 18 to 81. Hence, the school-age population also represents an adult group, but comparatively young adult group. The similarity in diversity between these two groups could be explained by the fact that both groups represent adults of different life stages, sharing similar environmental exposures, lifestyle factors, and

microbiome characteristics. However, a significant increase in diversity was observed in the senior group compared to both the adult and school-age groups. Indicating that the microbial communities in older individuals tend to be more diverse.

This pattern suggested that age-related changes in the human microbiome result in a more diverse abundance of microbial species as individual's age progresses. The increased diversity in seniors could be related to various factors, such as lifestyle changes, diet, or immune system alterations associated with aging. The significant difference between the school-age and senior groups might reflect development and age related shifts in the microbiome occurring throughout the lifespan.

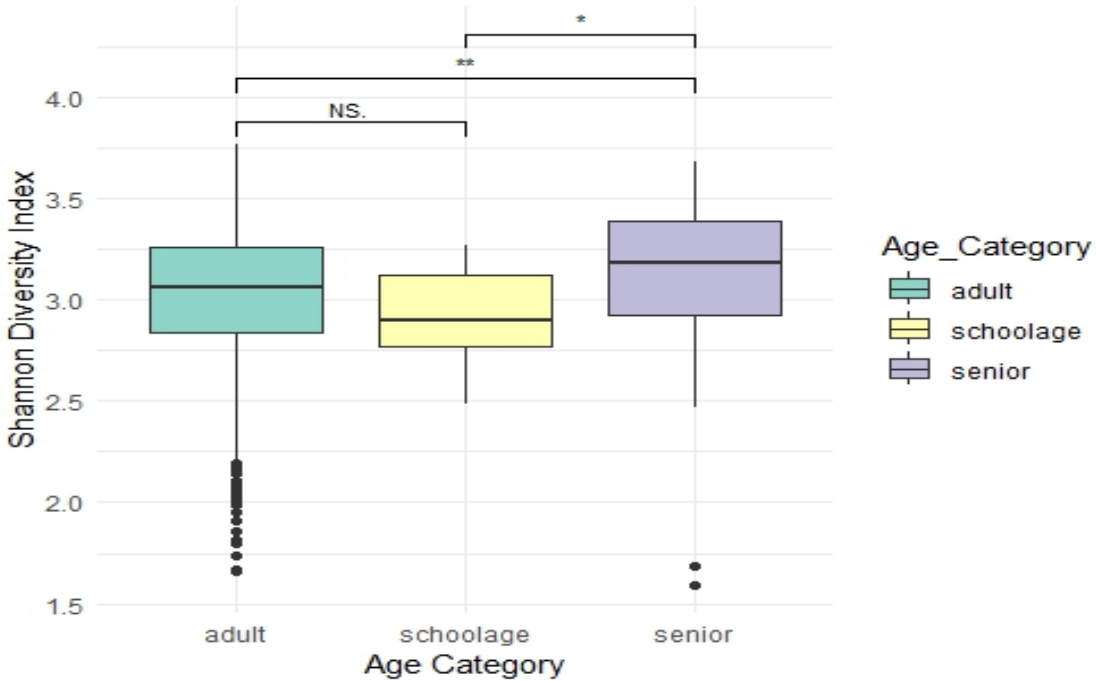


Figure 4.6: Boxplot showing Shannon diversity index values across different age categories, highlighting the variation in microbial diversity. The Shannon diversity index was computed for each sample based on species abundance data. Pairwise comparisons between age categories were performed using Dunn's test with Bonferroni adjustment. Significance levels are annotated with asterisks, indicating differences in microbial diversity between age groups (* $p < 0.05$, ** $p < 0.01$), while 'NS' denotes statistically not significant.

The PCoA plot 4.7 and PERMANOVA results 4.5 depicted interpretation of

beta diversity in the dataset. The PCoA plot showed a high degree of overlap and lack of distinct clustering between the samples from different age categories (adult, school-age, and senior). Suggesting that the microbial communities are relatively similar across these age groups. The wide spread of points indicated variability in microbial composition. The PERMANOVA test results also comply with the interpretation of figure 4.7. The R^2 value (0.00375) showed a fraction of the variation in species composition explained by age. However, the significant p-value (0.001) indicated that age still had an effect on microbial composition. While the statistical significance suggested subtle differences between age groups, but the microbial compositions were relatively similar across the age categories.

PERMANOVA results		
Sum of squares	**R^2**	**P Value**
0.83	0.00375	0.001

Table 4.5: PERMANOVA results based on age categories, showing the sum of squares, R^2, and p-value. The p-value indicates the statistical significance of the variation observed between age groups.

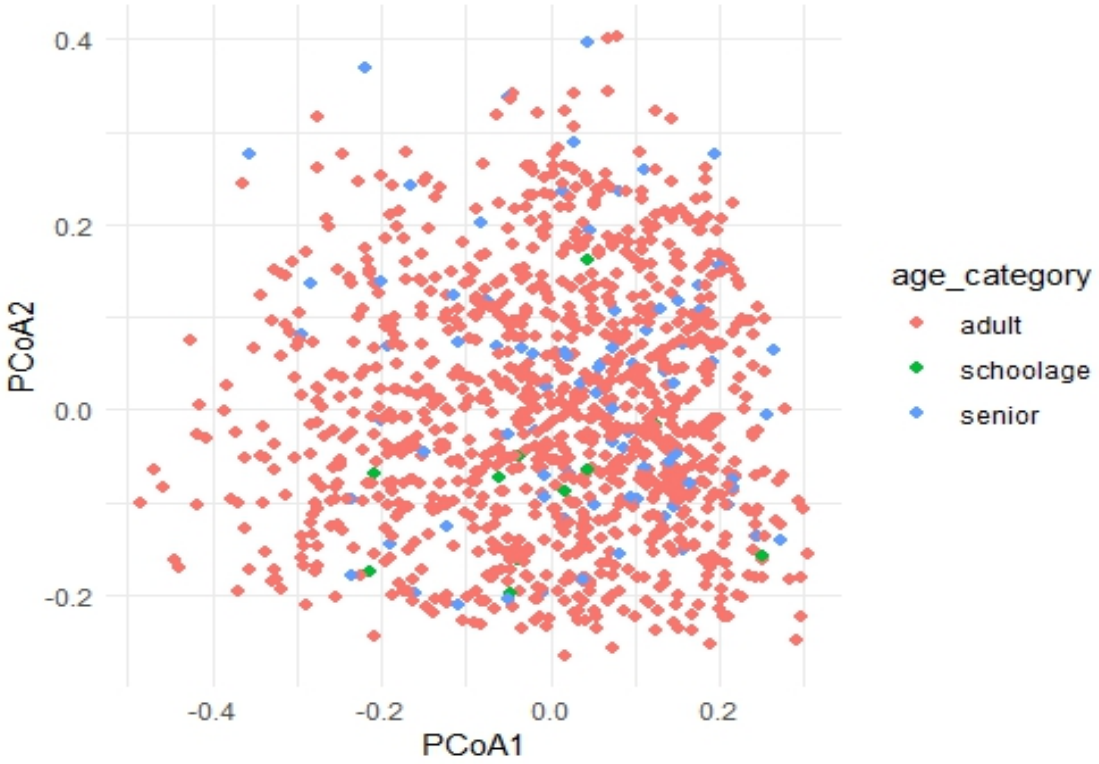


Figure 4.7: Beta diversity analysis illustrating community similarity in microbiome samples across different age categories.

4.5 Correlation Analysis

Figure 4.8 depicted a comparison of the correlation distributions of three microbiome data sets, Marker Abundance, Pathway Abundance, and Relative Abundance with the target variable, age. The boxplots illustrated the range and central tendency of correlation values for each dataset. Both Marker Abundance and Pathway Abundance exhibited weak but generally positive correlations with the target variable, suggesting a moderate and consistent association that may contain predictive information. In contrast, the Relative Abundance dataset showed a broader distribution of correlations centered around zero, including numerous negative values, indicating

a more variable and less stable relationship with the target variable.

Statistical significance markers between each pairwise comparison revealed that the differences in correlation distributions among the data sets were statistically significant. Specifically, the correlation distributions of Marker Abundance and Pathway Abundance were significantly higher and more consistent than those of Relative Abundance, suggesting that these data sets may provide more reliable predictive information.

These findings implied that the Marker and Pathway Abundance data sets may be more suitable for predictive modeling of the target variable due to their higher and more stable correlations. The Relative Abundance dataset, with its weaker and more variable correlations, may contribute less effectively to predictive accuracy. Therefore, prioritizing Marker and Pathway Abundance data sets in feature selection and model development could potentially enhance model performance in predicting age. This analysis highlighted the importance of selecting data sets with higher correlations to improve the reliability and interpretability of models in microbiome-based age prediction studies.

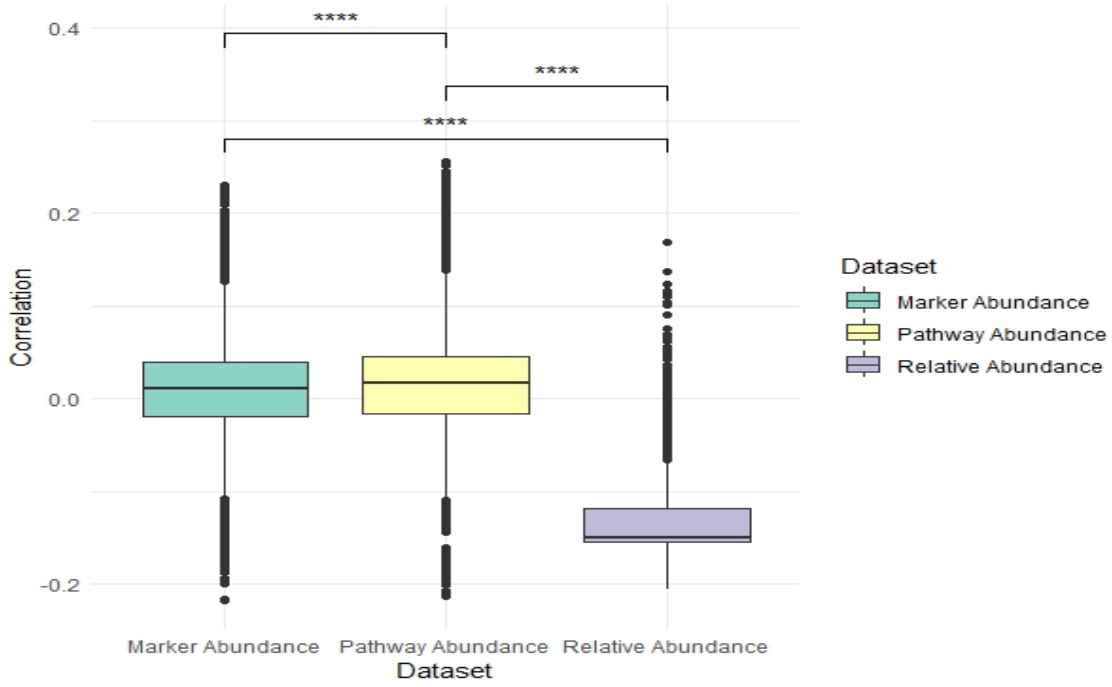


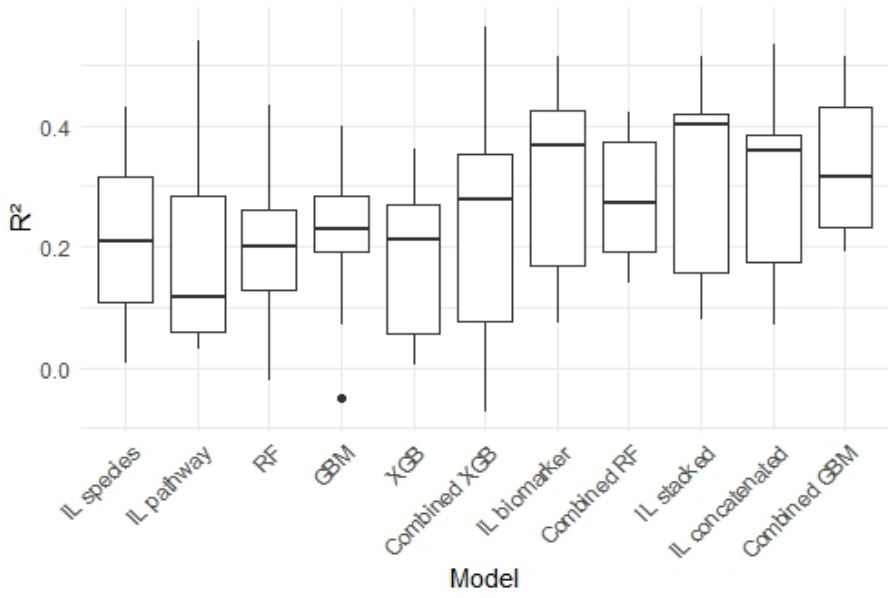
Figure 4.8: Boxplot comparing the distribution of feature correlations with age across three datasets: Relative Abundance, Marker Abundance, and Pathway Abundance. Each box represents the spread of correlation values within a dataset. Pairwise Wilcoxon rank-sum tests with Bonferroni correction were used to assess the significance of differences between datasets, with significance levels denoted by asterisks (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

4.6 Model Performance

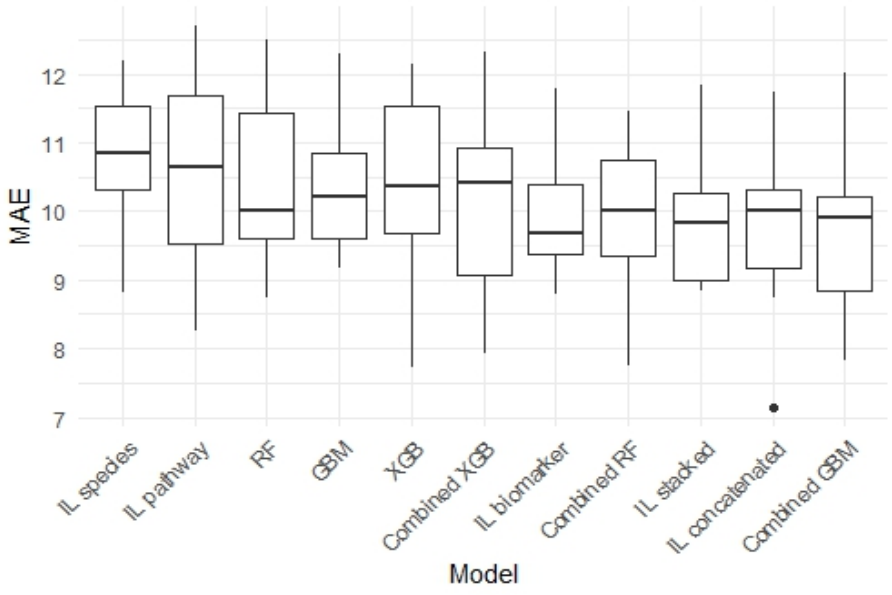
Figure 4.9(A) compared the performance of the machine learning models in both the single-omic and multi-omics data sets based on the R-squared (R^2) values. Models trained on multi-omics data sets (e.g., Combined RF, Combined XGB, Combined GBM, IL concatenated, and IL stacked) showed better performance, with higher R^2 values compared to those trained on individual data sets. Among these models, the Combined GBM model stands out with the highest mean R^2 value, showing com-

parable results to the stacked, concatenated configuration of the IntegratedLearner method. Additionally, the Combined GBM and Combined XGB models demonstrated smaller interquartile ranges (IQR) in R^2 values, indicating greater consistency in explaining the variance in the target variable. In contrast, models trained on individual data sets, such as GBM, RF, and XGB, exhibited greater variability in R^2, reflecting less stable performance. Notably, the model trained with the marker abundance dataset in the IntegratedLearner method outperformed some of the models trained on multi-omics data sets. Furthermore, the GBM model demonstrated enhanced performance, highlighting its effectiveness compared to RF and XGB when applied to single-omic configurations.

Additionally, Figure 4.9(B) presented a comparison of the models' performance based on MAE values. Overall, the findings indicated that the use of multi-omics dataset enhances the predictive accuracy of the models, as reflected in the performance metrics.



[A]



[B]

Figure 4.9: Performance comparison of machine learning models trained on single-omic (RF, XGB, GBM, IL species, IL pathway, and IL biomarker models) and multi-omics (Combined RF, Combined XGB, Combined GBM, IL concatenated, and IL stacked models) dataset evaluated on a test set using 10-fold cross-validation. Figure 4.9(A) shows the performance comparison of the models based on R-squared (R^2) values, while Figure 4.9(B) shows the comparison based on Mean Absolute Error (MAE) values.

The models were evaluated against a randomized baseline generated through permutation testing. In this approach, the target labels were randomly permuted, and the models were trained and tested on these permuted data sets. This baseline serves as a reference for assessing whether the models capture meaningful patterns or merely overfit to noise. All models, including single-omic (RF, XGB, GBM) and multi-omics (Combined RF, Combined XGB, Combined GBM), consistently outperformed the randomized baseline. This indicated that the models effectively captured meaningful patterns in the data, performing significantly better than random predictions.

Figure 4.10 depicted the relationship between actual and predicted ages using Gradient Boosting Machine (GBM), Random Forest (RF), and XGBoost models. The combined models aligned more closely with the diagonal line, indicating comparatively better predictive performance, while individual models showed variations in performance.

Moreover, figure 4.11 depicted the predictive performance of the IntegratedLearner models using different feature sets including biomarker, concatenated, pathway, species, and stacked models. The pathway and species models showed reasonable alignment with the diagonal line, suggesting moderate predictive performance. The concatenated and biomarker layers demonstrated good predictive performance, while the stacked model outperformed all of the models by achieving higher predictive performance, highlighting the effectiveness of integrating predictions from multiple layers in multi-omics analysis.

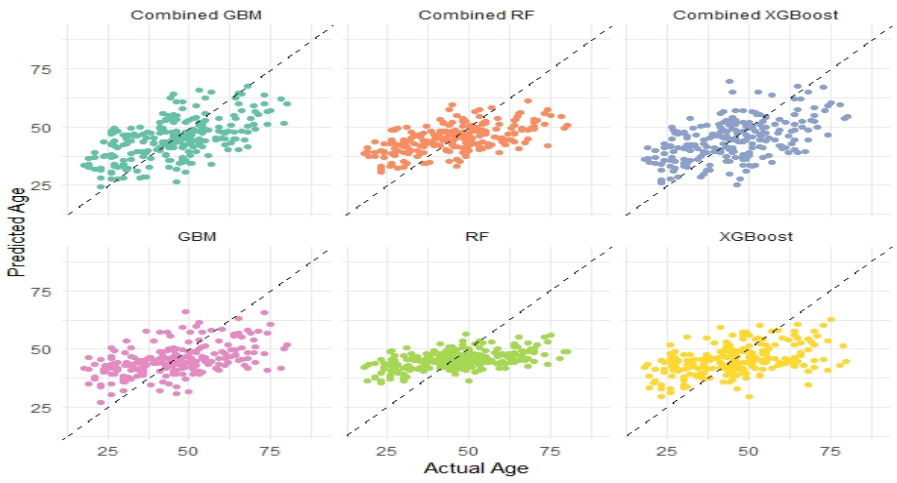


Figure 4.10: Jitter plots comparing actual and predicted ages for single-omic dataset trained models, GBM, RF, and XGBoost (bottom row) and multi-omics dataset models, Combined GBM, Combined RF, Combined XGBoost (top row). The diagonal line represents the standard prediction alignment.

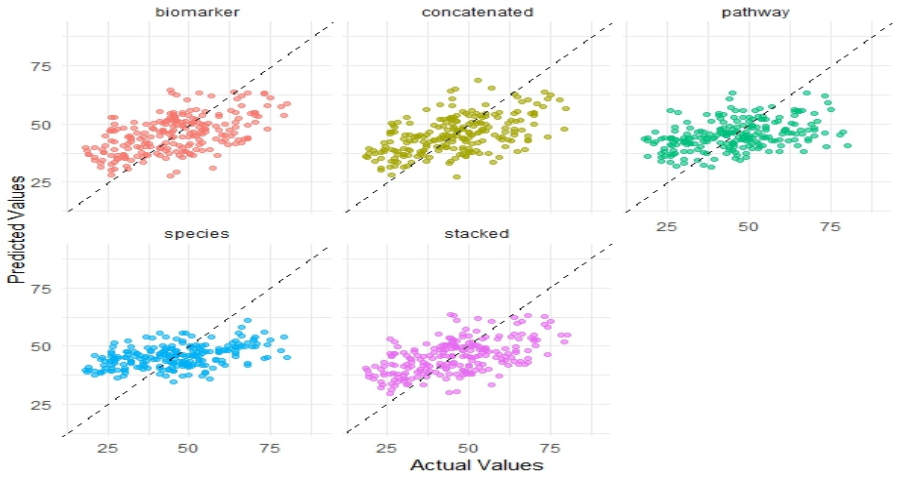


Figure 4.11: Jitter plots comparing actual and predicted ages across different layers (biomarker, concatenated, pathway, species, and stacked) of the IntegratedLearner method. The concatenated layer represents the multi-omics dataset, while the stacked layer serves as the meta-learner model, combining predictions from individual layers for improved accuracy. The diagonal line represents the standard prediction alignment.

Figure 4.12 evaluated the performance of the predictive models using R^2 values as a measure of their explanatory power. The scatter plot depicted the distribution of R^2 values across different models, including single-omic models (e.g., GBM, RF, IL species, IL pathway), and multi-omics models (e.g., Combined RF, IL stack, Combined GBM). Notably, models such as IL concatenated, Combined GBM, and IL stack achieved consistently higher R^2 values, suggesting these approaches were better suited for capturing the variance in the data. In contrast, models such as the IL species and the IL pathway exhibited lower R^2 values and greater variability, indicating limited predictive performance in this context.

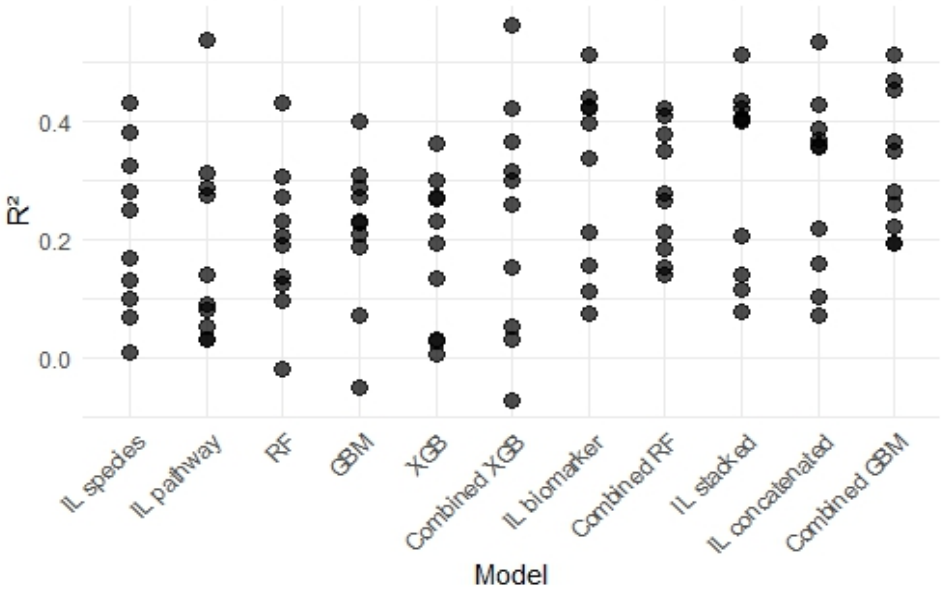


Figure 4.12: Scatter plot of R^2 values across Predictive Models. The plot compares the predictive performance of various models using R^2 values, with higher values indicating better model fit. Combined GBM and IL stack demonstrated superior performance compared to all the other single-omic and multi-omics model.

Table 4.6 and 4.7 presents the results of pairwise t-tests performed on the R^2 values of single-omics and multi-omics machine learning models. The table highlights p-values to assess statistical significance.

Pairwise t-test Results						
Group 1 / Group 2	**RF**	**XGB**	**GBM**	**RF (multi-omics)**	**XGB (multi-omics)**	**GBM (multi-omics)**
RF	-	0.7840	0.7709	0.1330	0.5874	**0.0258***
XGB	0.7840	-	0.5756	0.0794	0.4566	**0.0147***
GBM	0.7709	0.5756	-	0.2292	0.7474	**0.0489***
RF(multi-omics)	0.1330	0.0794	0.2292	-	0.5738	0.3257
XGB (multi-omics)	0.5874	0.4566	0.7474	0.5738	-	0.2274
GBM (multi-omics)	**0.0258***	**0.0147***	**0.0489***	0.3257	0.2274	-

Table 4.6: Pairwise t-test results comparing R^2 values across different machine learning models, including single-omics models (RF, XGB, GBM) and multi-omics models (RF, XGB, GBM with multi-omics integration). The table displays the p-values for each pairwise comparison, with (* $p < 0.05$) indicating statistically significant results.

4.7 Important Features

The tables presented the results of feature importance rankings derived from three machine learning models Random Forest, XGBoost, and Gradient Boosting Machine (GBM) for single-omic and multi-omics data sets. These results highlighted the features that were most influential in predicting the target variable across the data sets.

Table 4.8 showed the feature importance rankings using the Random Forest model, measured by the Increased Node Purity metric. In the single-omic dataset, the top feature was Clostridium disporicum with the highest Increased Node Purity of 5579.825, followed by Slackia isoflavoniconvertens and Firmicutes bacterium CAG 94. For the multi-omics dataset, pathway features such as PWY-5913 (5639.836)

Pairwise t-test Results					
Group 1 / Group 2	**RF (multi-omics)**	**XGB (multi-omics)**	**GBM (multi-omics)**	**Integrated Learner (stacked)**	**Integrated Learner (concatenated)**
RF(multi-omics)	-	0.2207	0.2954	0.1205	0.1619
XGB (multi-omics)	0.2207	-	0.5117	0.1205	0.1619
GBM (multi-omics)	0.2954	0.5117	-	0.5117	0.6495
Integrated Learner (stacked)	0.1205	0.1205	0.5117	-	0.8551
Integrated Learner (concatenated)	0.1619	0.1619	0.6495	0.8551	-

Table 4.7: Pairwise t-test results comparing R^2 values across different machine learning models, including single-omics models (RF, XGB, GBM) and multi-omics models (RF, XGB, GBM with multi-omics integration). The table displays the p-values for each pairwise comparison, with (* $p < 0.05$) indicating statistically significant results.

and PWY0-1586 (2168.797) dominated the rankings, reflecting the importance of functional pathways when multiple omics data are combined. The difference in top features between the single-omic and multi-omics data sets suggested that the integration of data types shifted the model's focus to metabolic pathways and specific protein annotations.

Table 4.9 showed feature importance rankings for XGBoost using three metrics Gain, Cover, and Frequency. For the single-omic dataset, Slackia isoflavoniconvertens showed the highest Gain (0.0449), indicating its significant contribution to improving model accuracy. Features like Firmicutes bacterium CAG 94 and

Clostridium disporicum were also highly ranked across all metrics. In the multi-omics dataset, pathways such as PWY-5913 and protein annotations like 418240 C6J849 ERS852478 02462 ranked highly in Gain, reflecting their strong influence on predictions. Similar to the Random Forest results, the feature ranking of the XGBoost model for the multi-omics dataset also highlighted a greater influence of pathway and marker abundance features compared to microbial taxa.

Table 4.10 provided feature rankings based on the GBM model, measured by Relative Influence. For the single-omic dataset, Slackia isoflavoniconvertens had the highest influence (5.5274), followed by Firmicutes bacterium CAG 94 and Clostridium disporicum. This estimation also aligned with the RF and XGBoost results, showing consistency in identifying important taxa. In the multi-omics dataset, pathway PWY-5913 was the top feature, with a Relative Influence of 2.7094. Other influential features included protein annotations such as 418240 C6J849 ERS852478 02462 and PWY0-1586.

In summary, all three models consistently identified key features for both single-omic and multi-omics data sets. This suggests that functional attributes, such as metabolic pathways and specific biomarkers, played more significant role in predicting the target variable, potentially reflecting their closer association with age. Moreover, Slackia isoflavoniconvertens, Firmicutes bacterium CAG 94, and Clostridium disporicum appeared as the top common features across all model results in the case of single-omic analysis.

Feature Ranking by Random Forest Model			
Single-omic		**Multi-omics**	
Feature	**Increased Node Purity**	**Feature**	**Increased Node Purity**
Clostridium disporicum	5579.825	PWY-5913	5639.836
Slackia isoflavoniconvertens	5241.173	PWY0-1586	2168.797
Firmicutes bacterium CAG 94	4888.641	28116 C3QQN3 CUY 4207	2117.562
Bifidobacterium adolescentis	4012.824	1261 D3MR67 SAMN05660467 01650	2109.406
Bacteroides ovatus	3409.038	2108523 A0A2P2F5G5 LAWASA 1244	1985.166
Escherichia coli	3302.729	FERMENTATION-PWY	1891.353
Roseburia faecis	3276.720	418240 C6J849 ERS852478 02462	1881.355
Bacteroides caccae	3115.153	PWY-6168	1872.308
Adlercreutzia equolifaciens	2946.564	Bilophila wadsworthia	1796.306
Agathobaculum butyriciproducens	2714.370	78257 A0A087CUE0 BSAE 1156	1762.814

Table 4.8: Feature ranking based on increased node purity using the Random Forest model for both single-omic and multi-omics datasets. The table lists the top features from each dataset, with their corresponding node purity values. The single-omic dataset includes individual microbiome species, while the multi-omics dataset features pathways and other related biomarkers.

Feature Ranking by XGBoost Model (single-omic dataset)			
Feature	**Gain**	**Cover**	**Frequency**
Slackia isoflavoniconvertens	0.0449	0.0374	0.0240
Firmicutes bacterium CAG 94	0.0437	0.0261	0.0254
Clostridium disporicum	0.0374	0.0382	0.0225
Ruminococcus torques	0.0323	0.0392	0.0314
Adlercreutzia equolifaciens	0.0323	0.0251	0.0210
Bacteroides ovatus	0.0305	0.0169	0.0225
Bifidobacterium adolescentis	0.0305	0.0239	0.0210
Bifidobacterium bifidum	0.0261	0.0219	0.0165
Agathobaculum butyriciproducens	0.0239	0.0232	0.0150
Roseburia faecis	0.0236	0.0195	0.0150
Feature Ranking by XGBoost Model (muti-omics dataset)			
Feature	**Gain**	**Cover**	**Frequency**
PWY-5913	0.0314	0.0157	0.0102
418240 C6J849 ERS852478 02462	0.0263	0.0184	0.0139
PWY-6385	0.0202	0.0034	0.0015
Bilophila wadsworthia	0.0173	0.0136	0.0117
28116 C3QQN3 CUY 4207	0.0172	0.0095	0.0066
2108523 A0A2P2F5G5 LAWASA 1244	0.0167	0.0213	0.0154
Streptococcus gordonii	0.0161	0.0063	0.0029
1262989 R6ZBG1 BN815 00721	0.0160	0.0058	0.0029
PWY0-1586	0.0160	0.0137	0.0095
1261 D3MR67 SAMN05660467 01650	0.0145	0.0170	0.0154

Table 4.9: Feature ranking based on the XGBoost model for both single-omic and multi-omics datasets. The table lists the top features from each dataset along with their corresponding gain, cover, and frequency values. Gain represents the importance of the feature in terms of the improvement it brings to the model, cover indicates the relative coverage of the feature in the dataset, and frequency shows how often the feature was used across all trees in the model.

Feature Ranking by GBM Model			
Single-omic		**Multi-omics**	
Feature	**Relative Influence**	**Feature**	**Relative Influence**
Slackia isoflavoniconvertens	5.5274	PWY-5913	2.7094
Firmicutes bacterium CAG 94	5.3900	418240 C6J849 ERS852478 02462	2.3130
Clostridium disporicum	5.2963	28116 C3QQN3 CUY 4207	1.6770
Agathobaculum butyriciproducens	3.9971	Bilophila wadsworthia	1.6323
Bacteroides ovatus	3.9785	2108523 A0A2P2F5G5 LAWASA 1244	1.6322
Roseburia faecis	3.7288	1261 D3MR67 SAMN05660467 01650	1.5623
Escherichia coli	3.6968	PWY0-1586	1.4915
Bifidobacterium adolescentis	3.4419	78257 A0A087CUE0 BSAE 1156	1.4451
Adlercreutzia equolifaciens	3.3813	PWY-7383	1.4295
Ruminococcus torques	3.1246	PWY-6168	1.3394

Table 4.10: Feature ranking based on the GBM model for both single-omic and multi-omics datasets. The table presents the top features from each dataset along with their corresponding relative influence values. Relative influence measures the importance of each feature in predicting the target variable, with higher values indicating greater importance. The table compares single-omic and multi-omics datasets, showing the most influential features in both contexts.

5 Discussion

The goal of this thesis was to evaluate and compare the predictive performance of the tree-based machine learning models, such as Random Forest (RF), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGBoost) using both the single-omic and multi-omics data sets, along with the multi-omics specialized IntegratedLearner method, for age prediction using the LifeLines DEEP dataset. This study further investigated whether single-omic dataset or multi-omics integration provided superior predictive performance, with particular emphasis on the capabilities of the IntegratedLearner models for multi-omics data analysis. All the models trained on multi-omics data sets, including RF, GBM, XGBoost, and the IntegratedLearner model, outperformed those trained on single-omics data sets.This suggested that the multi-omics integration enhanced the predictive performance of the models. While, the performance of RF, GBM, and XGBoost trained with multi-omics data was comparable with the IntegratedLearner model.

The LifeLines DEEP dataset posed several challenges for prediction due to the inherent characteristics of microbiome data. Across all data sets, the features exhibited right-skewed distributions, with most values concentrated near zero and a few outliers displayed higher abundances. This pattern reflects the compositional and sparse nature of microbiome data, where certain taxa, markers, or pathways are present in trace amounts, while others dominated only in specific samples. The relative abundance dataset, in particular, highlighted these challenges, as its com-

positional structure constrained the data, making it difficult for models to capture meaningful patterns or associations with age. Additionally, the diversity of microbial communities, measured by the number of distinct taxa, may influence the predictive power of the models. A higher microbial community diversity often correlates with richer biological information, potentially leading to more accurate predictions in certain contexts [63]. In this study, the Shannon diversity index revealed that microbial diversity increased with age, while the senior age group exhibited the highest diversity compared to the adult and school-age groups. However, the Beta diversity and PERMANOVA test indicated that there were subtle differences in microbial composition across age groups, the microbial communities were largely similar across these categories. Hence, the limited diversity observed in the LifeLines DEEP study may partially explain the challenges in accurately predicting age, underscoring the importance of microbial diversity in such analyses.

Microbiome research often emphasizes taxonomic and functional profiling of microbial communities, with functional profiling generally considered more effective for understanding human-microbe interactions [64]. In this study, the correlation analysis revealed that species relative abundance, representing taxonomic data, exhibited a nominal association with age. Conversely, functional data sets, specifically pathway abundance and marker abundance, showed stronger correlations with age. Functional data, which represents the biochemical activities and metabolic pathways of microbes, are inherently more stable across microbial communities, as multiple microbial species can perform similar functions. This stability enhances their robustness and reliability for predictive modeling, while taxonomic data often display significant variability and reduced consistency between different data sets or environments [64].

This disparity was further reflected in the predictive performance of the models developed in this study. Models trained exclusively on the relative abundance

dataset demonstrated lower predictive accuracy compared to those utilizing the functional data sets. However, integrating functional data sets (pathway abundance and marker abundance) with the relative abundance dataset to create a multi-omics dataset significantly enhanced model performance. Moreover, the results obtained using the IntegratedLearner method further validated the advantages of multi-omics integration for improving predictive performance.

The predictive performance of the tree-based models varied depending on the type of data set. For models trained on the multi-omics dataset, GBM achieved the highest performance (R^2 = 0.330, MAE = 9.66; see Table A.1), followed by RF (R^2 = 0.279, MAE = 9.92; see Table A.1) and XGBoost (R^2 = 0.239, MAE = 10.2; see Table A.1). A similar trend was observed for single-omic data sets, where GBM (R^2 = 0.214, MAE = 10.4; see Table A.1) consistently outperformed RF (R^2 = 0.198, MAE = 10.5; see Table A.1) and XGBoost (R^2 = 0.182, MAE = 10.4; see Table A.1). These results highlighted the consistent strength of GBM in predictive modeling, particularly when leveraging multi-omics data. Notably, the GBM model trained on the multi-omics dataset significantly outperformed (p-value < 0.05) the models trained on single-omic data sets, including GBM, RF, and XGBoost. This distinguished performance of the GBM model can be attributed to its ability to capture complex, non-linear relationships, iterative boosting that reduces overfitting, and effective feature prioritization, making it particularly well-suited for high-dimensional and sparse microbiome data [65] [66]. RF showed strong performance, especially when compared to XGBoost model. Its strength lies in its ability to handle high-dimensional data and its straightforward approach to hyperparameter tuning [67]. Similar studies on microbiome analysis have shown that RF can outperform XGBoost, although XGBoost is a more complex model [68].

Furthermore, the IntegratedLearner model evaluated individual omics layers, including Relative Abundance, Pathway Abundance, and Marker Abundance, while

leveraging their integration through a stacked layer as well as their combined representation. Among these layers, the stacked layer achieved the highest predictive performance (R^2 = 0.311, MAE = 9.87; see Table A.1). This result highlighted that training separate models for each data type and subsequently combining their outputs through a second layer effectively enhances model performance. Moreover, the concatenated layer (R^2 = 0.299, MAE = 9.76; see Table A.1) and the biomarker layer (R^2 = 0.308, MAE = 9.89; see Table A.1) demonstrated comparable performance, although both were outperformed by the stacked layer.

Lastly, the feature ranking analysis of the models revealed distinct patterns between single-omic and multi-omics data sets. For single-omic data sets, microbial species such as Clostridium disporicum, Slackia isoflavoniconvertens, and Firmicutes bacterium CAG 94 consistently ranked among the top features, highlighting their strong association with age prediction. In contrast, multi-omics data sets identified metabolic pathways and specific markers as critical predictors, reflecting the added complementary information provided by multi-omics integration. Notably, pathways such as PWY-5913 and PWY0-1586 emerged as significant contributors, emphasizing their key roles in age prediction. Beyond pathways, microbial taxa and specific markers also demonstrated substantial importance. For example, Bilophila wadsworthia, a microbial species, consistently appeared among the top-ranked features, underscoring its relevance in predictive models. Similarly, markers such as "28116 C3QQN3 CUY 4207" were repeatedly ranked among the top features, indicating a strong predictive significance.

The findings of this study highlighted the diverse contributions of taxa, pathways, and markers to age prediction, emphasizing the importance of multi-omics integration in capturing complex patterns. Furthermore, the incorporation of functional data sets, which provide a more stable and robust representation of microbial activities, improves predictive modeling and offers a deeper understanding of the

relationship between the microbiome and age.

6 Conclusion

This research aimed to predict age using the LifeLines DEEP dataset, incorporating Relative Abundance, Marker Abundance, and Pathway Abundance through machine learning and statistical methods. Pathway abundance showed the strongest correlation with age, highlighting its ability to capture age-related biological processes more effectively than other data sets. Diversity and indicator species analyses identified taxa linked to specific age categories, emphasizing their biological relevance and utility for predictive modeling.

The machine learning models demonstrated varying levels of predictive capability. While, GBM emerged as the most effective model, achieving superior accuracy and robustness, particularly when trained with the multi-omics dataset. While the IntegratedLearner stacked model achieved performance comparable to the GBM model (trained with the multi-omics dataset). Moreover, the result of the benchmark comparisons with a random baseline confirmed the statistical significance of the models, validating the feasibility of age prediction using microbiome data.

Despite the contributions, the study faced notable challenges. High sparsity and zero-inflated data sets posed significant hurdles for model training and prediction accuracy. Additionally, low correlations between certain features and the target variable, combined with limited diversity in age categories, constrained the predictive power of the data sets and impacted the generalizability of the findings. These limitations suggested avenues for improvement in future research. Expanding the

dataset quality and diversity will enhance generalizability, while adopting advanced techniques, such as deep learning with optimized hyperparameter tuning methods (e.g., grid search or Bayesian optimization), could further refine feature extraction and improve predictive accuracy.

In conclusion, this research demonstrates the potential of microbiome data for age prediction through machine learning and data analysis, providing a robust framework for this kind of study. By addressing current limitations and exploring innovative approaches, advancements in microbiome research can open the door to novel applications in personalized medicine, aging science, and beyond.

References

[1] S. C. Williams, *Our bacteria are more personal than we thought, stanford medicine-led study shows*, Stanford Medicine News Center, (2024) [Online]. Accessed: (4 December, 2024).

[2] *The evolution of bioinformatics*, BioStrand, (2022) [Online]. Available: `https://blog.biostrand.ai/the-evolution-of-bioinformatics`, Accessed: (4 December 2024).

[3] R. Turner, *Essentials of microbiology*. Scientific e-Resources, 2018, p. 65.

[4] *NIH human microbiome project defines normal bacterial makeup of the body*, National Institutes of Health, (2012)[Online]. Accessed: (22 February, 2024).

[5] H. Gest, "The discovery of microorganisms by robert hooke and antoni van leeuwenhoek, fellows of the royal society", *Notes and records of the Royal Society of London*, vol. 58, no. 2, pp. 187–201, 2004.

[6] J. Goins, *Microbiomes: An origin story*, American Society for Microbiology, (2019) [Online]. Accessed: (22 February, 2024).

[7] L. K. Ursell, J. L. Metcalf, L. W. Parfrey, and R. Knight, "Defining the human microbiome", *Nutrition reviews*, vol. 70, no. suppl_1, S38–S44, 2012.

[8] G. A. Ogunrinola, J. O. Oyewale, O. O. Oshamika, and G. I. Olasehinde, "The human microbiome and its impacts on health", *International journal of microbiology*, vol. 2020, no. 1, p. 8 045 646, 2020.

[9] M. H. Mohajeri, R. J. Brummer, R. A. Rastall, *et al.*, "The role of the microbiome for human health: From basic science to clinical applications", *European journal of nutrition*, vol. 57, pp. 1–14, 2018.

[10] M. J. Bull and N. T. Plummer, "Part 1: The human gut microbiome in health and disease", *Integrative Medicine (Encinitas)*, vol. 13, no. 6, pp. 17–22, 2014.

[11] Z. Y. Kho and S. K. Lal, "The human gut microbiome–a potential controller of wellness and disease", *Frontiers in microbiology*, vol. 9, p. 1835, 2018.

[12] E. Ragonnaud and A. Biragyn, "Gut microbiota as the key controllers of "healthy" aging of elderly people", *Immunity & Ageing*, vol. 18, pp. 1–11, 2021.

[13] R. K. Sah, A. Nandan, S. Prashant, *et al.*, "Decoding the role of the gut microbiome in gut-brain axis, stress-resilience, or stress-susceptibility: A review", *Asian Journal of Psychiatry*, p. 103 861, 2023.

[14] M. Carabotti, A. Scirocco, M. A. Maselli, and C. Severi, "The gut-brain axis: Interactions between enteric microbiota, central and enteric nervous systems", *Annals of gastroenterology: quarterly publication of the Hellenic Society of Gastroenterology*, vol. 28, no. 2, p. 203, 2015.

[15] T. Newman, *All about the central nervous system*, MedicalNewsToday, (2023)[Online]. Available:`https://www.medicalnewstoday.com/articles/307076?c=819621198226`, Accessed: (24 April, 2024).

[16] V. Philip and P. Bercik, "Gastrointestinal microbiota and the neural system", in *The Microbiota in Gastrointestinal Pathophysiology*, Elsevier, 2017, pp. 243–247.

[17] K. Hou, Z.-X. Wu, X.-Y. Chen, *et al.*, "Microbiota in health and diseases", *Signal transduction and targeted therapy*, vol. 7, no. 1, pp. 1–28, 2022.

[18] B. Madhogaria, P. Bhowmik, and A. Kundu, "Correlation between human gut microbiome and diseases", *Infectious Medicine*, vol. 1, no. 3, pp. 180–191, 2022.

[19] S. P. Wiertsema, J. van Bergenhenegouwen, J. Garssen, and L. M. Knippels, "The interplay between the gut microbiome and the immune system in the context of infectious diseases throughout life and the role of nutrition in optimizing treatment strategies", *Nutrients*, vol. 13, no. 3, p. 886, 2021.

[20] T. Choden and N. A. Cohen, "The gut microbiome and the immune system", *Exploration of Medicine*, vol. 3, no. 3, pp. 219–233, 2022.

[21] N. Akbar, N. A. Khan, J. S. Muhammad, and R. Siddiqui, "The role of gut microbiome in cancer genesis and cancer prevention", *Health Sciences Review*, vol. 2, p. 100 010, 2022.

[22] A. Nogal, A. M. Valdes, and C. Menni, "The role of short-chain fatty acids in the interplay between gut microbiota and diet in cardio-metabolic health", *Gut microbes*, vol. 13, no. 1, p. 1 897 212, 2021.

[23] H. Wang, Y. Chen, L. Feng, *et al.*, "A gut aging clock using microbiome multi-view profiles is associated with health and frail risk", *Gut Microbes*, vol. 16, no. 1, p. 2 297 852, 2024.

[24] S. Huang, N. Haiminen, A.-P. Carrieri, *et al.*, "Human skin, oral, and gut microbiomes predict chronological age", *Msystems*, vol. 5, no. 1, pp. 10–1128, 2020.

[25] E. Bradley and J. Haran, "The human gut microbiome and aging", *Gut Microbes*, vol. 16, no. 1, p. 2 359 677, 2024.

[26] V. D. Badal, E. D. Vaccariello, E. R. Murray, *et al.*, "The gut microbiome, aging, and longevity: A systematic review", *Nutrients*, vol. 12, no. 12, p. 3759, 2020.

[27] E. Sepp, I. Smidt, T. Rööp, *et al.*, "Comparative analysis of gut microbiota in centenarians and young people: Impact of eating habits and childhood liv-

ing environment", *Frontiers in cellular and infection microbiology*, vol. 12, p. 851 404, 2022.

[28] N. Aggarwal, S. Kitano, G. R. Y. Puah, S. Kittelmann, I. Y. Hwang, and M. W. Chang, "Microbiome and human health: Current understanding, engineering, and enabling technologies", *Chemical reviews*, vol. 123, no. 1, pp. 31–72, 2022.

[29] N. Bosco and M. Noti, "The aging gut microbiome and its impact on host immunity", *Genes & Immunity*, vol. 22, no. 5, pp. 289–303, 2021.

[30] T. Wilmanski, C. Diener, N. Rappaport, *et al.*, "Gut microbiome pattern reflects healthy ageing and predicts survival in humans", *Nature metabolism*, vol. 3, no. 2, pp. 274–286, 2021.

[31] R. K. Weersma, A. Zhernakova, and J. Fu, "Interaction between drugs and the gut microbiome", *Gut*, vol. 69, no. 8, pp. 1510–1519, 2020.

[32] Y. Chen, H. Wang, W. Lu, *et al.*, "Human gut microbiome aging clocks based on taxonomic and functional signatures through multi-view learning", *Gut Microbes*, vol. 14, no. 1, p. 2 025 016, 2022.

[33] S.-H. Seo, C.-S. Na, S.-E. Park, *et al.*, "Machine learning model for predicting age in healthy individuals using age-related gut microbes and urine metabolites", *Gut Microbes*, vol. 15, no. 1, p. 2 226 915, 2023.

[34] F. Galkin, A. Aliper, E. Putin, I. Kuznetsov, V. N. Gladyshev, and A. Zhavoronkov, "Human microbiome aging clocks based on deep learning and tandem of permutation feature importance and accumulated local effects", *BioRxiv*, p. 507 780, 2018.

[35] H. Mallick, A. Porwal, S. Saha, P. Basak, V. Svetnik, and E. Paul, "An integrated bayesian framework for multi-omics prediction and classification", *Statistics in Medicine*, vol. 43, no. 5, pp. 983–1002, 2024.

[36] R. C. Gentleman, V. J. Carey, D. M. Bates, *et al.*, "Bioconductor: Open software development for computational biology and bioinformatics", *Genome biology*, vol. 5, pp. 1–16, 2004.

[37] T. Borman, F. G. Ernst, S. A. Shetty, and L. Lahti, *Mia: Microbiome analysis*, R package version 1.15.6, 2024. [Online]. Available: `https://github.com/microbiome/mia`.

[38] P. J. McMurdie and S. Holmes, "Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data", *PLoS ONE*, vol. 8, no. 4, e61217, 2013.

[39] E. Pasolli, L. Schiffer, P. Manghi, *et al.*, "Accessible, curated metagenomic data through ExperimentHub", en, *Nat. Methods*, vol. 14, no. 11, pp. 1023–1024, Oct. 2017, ISSN: 1548-7091, 1548-7105. DOI: `10.1038/nmeth.4468`.

[40] M. S. Matchado, M. Rühlemann, S. Reitmeier, *et al.*, "On the limits of 16s rrna gene-based metagenome prediction and functional profiling", *Microbial Genomics*, vol. 10, no. 2, p. 001 203, 2024.

[41] E. F. Tigchelaar, A. Zhernakova, J. A. Dekens, *et al.*, "Cohort profile: Lifelines deep, a prospective, general population cohort study in the northern netherlands: Study design and baseline characteristics", *BMJ open*, vol. 5, no. 8, e006772, 2015.

[42] A. Zhernakova, A. Kurilshikov, M. J. Bonder, *et al.*, "Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity", *Science*, vol. 352, no. 6285, pp. 565–569, 2016.

[43] C. Moreno, I. Zuria, M. García-Zenteno, *et al.*, "Trends in the measurement of alpha diversity in the last two decades", *Interciencia*, vol. 31, no. 1, pp. 67–71, 2006.

[44] A. K. Thukral, "A review on measurement of alpha diversity in biology.", *Agricultural Research Journal*, vol. 54, no. 1, 2017.

[45] J. D. Bakker, "Common distance measures", *Applied Multivariate Statistics in R*, 2024.

[46] D. Lynch, *Beta diversity: An introduction to beta diversity, and some of the common ways to measure it*, ONECODEX, (2021) [Online]. Available: `https://docs.onecodex.com/en/articles/4150649-beta-diversity`, Accessed: (14 November, 2024).

[47] M. Dufrêne and P. Legendre, "Species assemblages and indicator species: The need for a flexible asymmetrical approach", *Ecological monographs*, vol. 67, no. 3, pp. 345–366, 1997.

[48] M. De Cáceres and P. Legendre, "Associations between species and groups of sites: Indices and statistical inference", *Ecology*, vol. 90, pp. 3566–3574, 2009. DOI: `10.1890/08-1823.1`.

[49] J. D. Bakker, "Isa", *Applied Multivariate Statistics in R*, 2024.

[50] D. R. Cox, "The regression analysis of binary sequences", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, 1958.

[51] T. K. Ho, "Random decision forests", in *Proceedings of 3rd international conference on document analysis and recognition*, IEEE, vol. 1, 1995, pp. 278–282.

[52] C. Cortes and V. Vapnik, "Support-vector networks", *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[53] R. Tibshirani, "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[54] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, San Francisco, California, USA: ACM, 2016, pp. 785–794, ISBN: 978-1-4503-4232-2.

[55] J. H. Friedman, "Greedy function approximation: A gradient boosting machine", *Annals of statistics*, pp. 1189–1232, 2001.

[56] Kuhn and Max, "Building predictive models in r using the caret package", *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008.

[57] J. Brownlee, *Caret r package for applied predictive modeling*, Machine Learning Mastery, (2014) [Online]. Available: `https://machinelearningmastery.com/caret-r-package-for-applied-predictive-modeling`, Accessed: (18 August 2024).

[58] H. A. Chipman, E. I. George, and R. E. McCulloch, "Bart: Bayesian additive regression trees", *Duke University Press*, 2010.

[59] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome datasets are compositional: And this is not optional", *Frontiers in microbiology*, vol. 8, p. 2224, 2017.

[60] L. Lahti and S. Shetty, *Microbiome r package*, 2012-2019.

[61] Y.-H. Zhou and P. Gallins, "A review and tutorial of machine learning methods for microbiome host trait prediction", *Frontiers in genetics*, vol. 10, p. 579, 2019.

[62] M. Slawski and M. Hein, "Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization", *Duke University Press*, 2013.

[63] L. Wu, X.-W. Wang, Z. Tao, *et al.*, "Data-driven prediction of colonization outcomes for complex microbial communities", *Nature Communications*, vol. 15, no. 1, p. 2406, 2024.

[64] M. G. Langille, "Exploring linkages between taxonomic and functional profiles of the human microbiome", *MSystems*, vol. 3, no. 2, pp. 10–1128, 2018.

[65] P. Florek and A. Zagdański, "Benchmarking state-of-the-art gradient boosting algorithms for classification", *arXiv preprint arXiv:2305.17094*, 2023.

[66] B. C. Perez, M. C. Bink, K. L. Svenson, G. A. Churchill, and M. P. Calus, "Prediction performance of linear models and gradient boosting machine on complex phenotypes in outbred mice", *G3*, vol. 12, no. 4, jkac039, 2022.

[67] D. Ghosh and J. Cabrera, "Enriched random forest for high dimensional genomic data", *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 19, no. 5, pp. 2817–2828, 2021.

[68] B. D. Topçuoğlu, N. A. Lesniak, M. T. Ruffin IV, J. Wiens, and P. D. Schloss, "A framework for effective application of machine learning to microbiome-based classification problems", *MBio*, vol. 11, no. 3, pp. 10–1128, 2020.

Appendix A Appendix

Model Performance Metrics		
Model	R^2	**MAE**
GBM (multi-omics)	0.330	9.66
IntegratedLearner (stacked)	0.311	9.87
IntegratedLearner (biomarker)	0.308	9.89
IntegratedLearner (concatenated)	0.299	9.76
RF (multi-omics)	0.279	9.92
XGB (multi-omics)	0.239	10.2
IntegratedLearner (species)	0.215	10.8
GBM	0.214	10.4
RF	0.198	10.5
IntegratedLearner (pathway)	0.184	10.65
XGB	0.182	10.4

Table A.1: Model performance metrics for predicting age, sorted by R^2 values in descending order.

Appendix B Appendix

The code for this thesis can be found on GitHub at this link.